UC Berkeley
UC Berkeley Electronic Theses and Dissertations

Title
Computational approaches to understanding the genetic architecture of complex traits

Permalink
https://escholarship.org/uc/item/86q5k57j

Author
Brown, Brielin C.

Publication Date
2016

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Computational approaches to understanding the genetic architecture of complex traits

by

Brielin Chase Brown

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Lior Pachter, Co-chair
Assistant Professor Noah Zaitlen, Co-chair
Professor Satish Rao
Professor Lisa Barcellos

Fall 2016

Computational approaches to understanding the genetic architecture of complex traits

Copyright 2016
by
Brielin Chase Brown

Abstract

Computational approaches to understanding the genetic architecture of complex traits

by

Brielin Chase Brown

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor Lior Pachter, Co-chair

Assistant Professor Noah Zaitlen, Co-chair

Advances in DNA sequencing technology have resulted in the ability to generate genetic data at costs unimaginable even ten years ago. This has resulted in a tremendous amount of data, with large studies providing genotypes of hundreds of thousands of individuals at millions of genetic locations. This rapid increase in the scale of genetic data necessitates the development of computational methods that can analyze this data rapidly without sacrificing statistical rigor.

The low cost of DNA sequencing also provides an opportunity to tailor medical care to an individuals unique genetic signature. However, this type of *precision* medicine is limited by our understanding of how genetic variation shapes disease. Our understanding of so-called *complex* diseases is particularly poor, and most identified variants explain only a tiny fraction of the variance in the disease that is expected to be due to genetics. This is further complicated by the fact that most studies of complex disease go directly from genotype to phenotype, ignoring the complex biological processes that take place in between.

Herein, we discuss several advances in the field of complex trait genetics. We begin with a review of computational and statistical methods for working with genotype and phenotype data, as well as a discussion of methods for analyzing RNA-seq data in effort to bridge the gap between genotype and phenotype. We then describe our methods for 1) improving power to detect common variants associated with disease, 2) determining the extent to which different world populations share similar disease genetics and 3) identifying genes which show differential expression between the two haplotypes of a single individual. Finally, we discuss opportunities for future investigation in this field.

To Claire

Contents

Contents **ii**

List of Figures **iii**

List of Tables **vi**

1 Introduction **1**
1.1 Complex traits and the problem of missing heritability 1
1.2 Statistical models for complex trait genetics 6
1.3 Gene expression as a genetic trait 10

2 Local joint testing improves power and identifies hidden heritability in association studies **17**
2.1 Introduction 17
2.2 Methods 19
2.3 Results 25
2.4 Discussion 33

3 Transethnic genetic correlations from summary statistics **39**
3.1 Introduction 39
3.2 Methods 41
3.3 Results 45
3.4 Discussion 53

4 Allele-specific transcript abundance estimation **57**
4.1 Introduction 57
4.2 Results 58
4.3 Methods 62
4.4 Discussion 65

5 Discussion **68**

Bibliography **70**

List of Figures

1.1 An example of a manhattan plot with simulated data. Each of the points is $-\log_{10}(p)$, where the p-value is determined by the χ^2-test statistic of association with the phenotype. The dashed line represents $-\log_{10}(5 \times 10^{-8})$, the threshold of statistical significance in most genome-wide association studies. 5

1.2 Distribution of the underlying liability in a case-control study, using the liability threshold model. In both cases, the population prevalence of the trait is $k = 6\%$ and there are $100,000$ individuals. The threshold at which an individual is considered a case is $\Phi^{-1}(1-k) = 1.55$. (A) With no ascertainment, the underlying liability has a normal distribution. (B) In most studies, there are many more cases than in the general population. In this example, there are $50,000$ cases and $50,000$ controls. 8

2.1 Comparison of the density of the maximum test statistic from local joint testing on the first 1000 SNPs of chromosome 1 in the WT controls between *Jester* and permutation approaches. In each case 100K samples were used. The significance level α corresponding to an FWER of 5% for *Jester* in this experiment was $\alpha_{jester} = 4.37\text{e}{-06}$ and for the permutation test was $\alpha_{permutation} = 4.66\text{e}{-06}$. For the purposes of this plot, marginal test statistics (Z values) and joint test statistics (χ^2_2) were transformed to χ^2_1. 23

2.2 (Left) Joint testing genome-wide shows a power loss for all correlation structures when only one SNP affects the trait. (Right) Joint testing genome-wide shows a substantial power gain for anti-correlated SNPs that both affect the trait. . . . 26

2.3 The heritability of liability due to genome-wide significant marginal associations (dark blue) plus additional heritability explained by genome-wide significant joint associations (light blue). Error bars correspond to the standard error of the heritability estimates. In all cases but the T2D GERA, p-values correspond to the liklihood ratio test of the linear mixed-model fit with both marginal and joint GRM against the linear-mixed model fit only with the marginal GRM. In the T2D GERA case, the p-value corresponds to a likelihood ratio test of the linear model fit will joint and marginal significant SNPs against the model fit with only marginally-significant SNPs. 31

2.4 (left) Density of the correlation between SNPs in pairs that are genome-wide significant at FWER 5% in the Wellcome Trust dataset, with all pairs of significant SNPs in light blue and just those SNPs discovered using `jester` in dark blue. (right) Density of the correlation between SNPs in the most significant pair for each gene containing an eQTL pair at FDR 5% in the gEUVADIS dataset, with all pairs from genes with significant eQTLs in light blue and just those eQTL's discovered using `jester` in dark blue. 35

2.5 Q-Q plots of the results of FaST-LMM on the WTCCC dataset. (A) Control-control analysis, 15.6% of sets have a p-value of less than 0.10. (B) Bipolar disorder, 18.4% of sets have a p-value of less than 0.10. (C) Coronary artery disease, 13.4% of sets have a p-value of less than 0.10. (D) Crohn's disease, 16.5% of sets have a p-value of less than 0.10. (E) Hypertension, 13.4% of sets have a p-value of less than 0.10. (F) Rheumatoid arthritis, 14.2% of sets have a p-value of less than 0.10. (G) Type-1 diabetes, 13.7% of sets have a p-value of less than 0.10. (H) Type-2 diabetes, 14.7% of sets have a p-value of less than 0.10. (I) Bipolar disorder with leave-one-chromosome-out GRM background kernel, 15.4% of sets have a p-value of less than 0.10. In all cases we used 100-SNP sets. In (A) through (H), we used a likelihood ratio test with 10 permutations per set and no background kernel. In (I), we used the `sc_davies` score test to improve speed with the background kernel present. In some cases, λ_{GC} is 0 because permutation tests result in more than half of the sets having a p-value of 1.0. 38

3.1 Bias and standard error of the heritability estimator in `popcorn` as the number of SNPs M and number of individuals N varies.. All simulations conducted using simulated phenotypes with $h_1^2 = 0.5$, $h_2^2 = 0.5$, $\rho_{gi,e} = 0.5$ and simulated European (EUR) and East Asian (EAS) genotypes generated with HapGen2. 42

3.2 True and estimated genetic impact and effect correlation. All simulations conducted with simulated EUR and EAS heritability of 0.5 using 4499 simulated EUR and 4836 simulated EAS individuals at 248,953 SNPs. 46

3.3 Bias and standard error as the proportion of causal variants is decreased from 1.0 (all variants causal, the infinitesimal model) to 0.0001 (one in ten thousand variants causal, or approximately 25 total causals). All simulations conducted using simulated phenotypes with $h_1^2 = 0.5$, $h_2^2 = 0.5$, $\rho_{gi,e} = 0.5$ and simulated European (EUR) and East Asian (EAS) genotypes generated with HapGen2. . 47

3.4 Kernel density estimate (KDE) comparison between `popcorn` and GCTA as heritability estimators. Density was computed using the scipy statistics package `gaussian_kde` function on the set of heritability estimates. 48

3.5 Kernel density estimate of genetic correlation comparison between `popcorn` and GCTA. Distribution was computed using a gaussian kde on the set of genetic correlation estimates.. 50

3.6 Genetic correlation as a function of heritability for gene expression. The mean and standard error of the genetic correlation of the set of genes with h12 and h22 exceeding threshold h in each analysis (y-axis) is plotted against h (x-axis). . . . 51

3.7 Null distribution of the conditional genetic correlation. Phenotypes were generated with heritability randomly sampled from the joint distribution of the gEUVADIS heritability estimates over randomly selected 4000 SNP regions from chromosome 1 of the true EUR and YRI genotypes and genetic correlation of 0. The mean and standard error of the genetic correlation of the set of genes with $\hat{h}_1^2$ and $\hat{h}_2^2$ exceeding threshold h in each analysis (y-axis) is plotted against h (x-axis) . 52
3.8 Comparison of `popcorn` and ldsc as heritability estimators as the number of SNPs and individuals in each study varies. All simulations conducted using simulated phenotypes with $h_1^2 = 0.5$, $h_2^2 = 0.5$, $\rho_{gi,e} = 0.5$ and simulated European (EUR) and East Asian (EAS) genotypes generated with HapGen2. 55

4.1 Distribution of heterozygous SNPs in gEUVADIS individuals. We replicate the finding of Castel et al [17] that most genes contain few hets, and observe that 95% of genes have 8 or fewer heterozygous SNPs. 59
4.2 KDE of the distribution of the "ground-truth" counts generated by our simulation. Due to the long tail of the distribution, we limit the x-axis for improved visualization. 60
4.3 True (x-axis) versus estimated (y-axis) counts for reference (blue) and alternate (green) alleles using (top) `ursa` with haplotypes, (middle) `ursa` without haplotypes and (bottom) `allelecounter`. In each case, we overlay the line of best fit from a linear regression of the estimated counts on the true counts. 61
4.4 Comparison of estimates of reference (blue) and alternate (green) counts using `ursa` with (x-axis) and without (y-axis) haplotype information. The estimates are highly concordant, with a correlation coefficient of $\rho = 0.995$ for the reference count and a correlation coefficient of $\rho = 0.994$ for the alternate count. 62
4.5 Violin plot of the error in the count estimate for `ursa` with and without haplotype information. 63
4.6 Simulation pipeline for generating personalized RNA-seq reads with allele-specific abundances. The hg19 reference FASTA and GTF are used to build the standard reference transcriptome, which is used in quantification of transcript abundances in one population with `kallisto`. These abundances are used to calculate parameters of NB distributions for the counts of each gene. These distributions are used to draw counts for the simulation, and fixed or random allele-specific effects can be added. The 1000 genomes VCFs are used by `ursa` to build personalized references. RSEM is used to build an error model for the simulator. Finally, the RSEM simulator is run independently for each haplotype with haplotype-specific counts to generate personalized RNA-seq reads, which are combined to erase haplotype of origin. 66

List of Tables

1.1 Complex traits are caused by hundreds or thousands of genetic variants and the environment, while mendelian traits are effected by a single genetic variant in a dominant or recessive pattern. Complex traits are the focus of this manuscript. 2

1.2 Type and number of various kinds of human genetic variation. Single nucleotide polymorphisms (SNPs) are the most common, making up about 95% of all variation. In each case, an example modification to the sequence GATTACA is provided. Note that there are many kinds of structural variation, and the example provided is a copy-number variant. 2

1.3 Examples of terms that can be included when modeling the relationship between genotype, environment and phenotype. For the purposes of this manuscript we will focus on additive genetic effects, while acknowledging the potential significance of other kinds of effects in later sections. The ellipsis indicates than in each case we model many additional effects. The notation $\mathbb{1}[C]$ is an indicator variable that the specified condition C holds. 3

1.4 Typical summary association data consists of SNP names (rsids), estimates of the effect size and stand error of that SNP, the reference and alternate alleles in the study, and the number of individuals with data at that SNP. 9

2.1 (Left) Total number of loci containing genome-wide significant SNPs discovered using standard marginal, local joint, SnipSnip (SS), and genome-wide imputation (imp) testing methods. (Right) Total number of genome-wide significant SNPs discovered using standard marginal, local joint, SnipSnip (SS) and genome-wide imputation (imp) testing methods. For our analysis of T1D and RA, we removed chromosome 6 because of the large effect HLA locus. 28

2.2 Loci that were not significant in the standard marginal approach but became significant using *Jester*. ρ indicates correlation of SNPs signed with respect to opposite effect direction. Results at these loci from imputation against 1000 genomes are also reported. P-values which are significant for a particular testing method are denoted by an asterisk. 28

2.3 Loci containing a marginally significant SNP at the 0.05 level after correction for genome-wide multiple testing (p ¡ 2.18e-7) 29

2.4 Loci significant in the WTCCC consortium at level 0.05 after correction for multiple testing of all SNPs with their 100 closest neighbors. A reference to the NHGRI study first reporting the association is provided. 30
2.5 Results of running SnipSnip on the WT disease cohort using the default parameters. All pairs with correlation above 0.8 were removed to filter false positives. 32
2.6 Replicated loci found in a GWAS on WTCCC variants imputed to 1000 genomes. The Comment column indicates whether the locus was detected in the marginal or 100-SNP joint method, and provides a reference to the earliest replecation for the loci not found in the standard marginal or local joint approaches 33
2.7 A sample of loci which appeared to be false positives in our dataset. In all cases, marginal signal was eliminated after replacing genotypes with those estimated from imputation against 1000 genomes. 34
2.8 Joint testing pairs of *cis*-SNPs improves the number of eQTL's detected at FDR 5% by 10.7%. Many of the new genes (row new) discovered using the joint test appear to be linkage masked (row LM), with correlation between the significant SNP pair of above 0.2. 35

3.1 Average heritability and genetic correlation over 1000 simulations with varying levels of ascertainment. All simulations contained exactly N cases and controls for a study prevalence of 0.5. Phenotypes were simulated with liability scale heritability of 0.3 for both phenotypes and genetic correlation of 0.3. 46
3.2 True and estimated values of the genetic impact and effect correlation in simulated EUR-like and EAS-like genotypes. Results are the average of 100 simulations with phenotype heritability of 0.5 in each population. 48
3.3 Heritability and genetic correlation of RA and T2D between EUR and EAS populations. EUR RA data contained 8,875 cases and 29,367 controls for a study prevalence of 0.23. EAS RA data contained 4,873 cases and 17,642 controls for a study prevalence of 0.22. RA disease prevalence was assumed to be 0.5% in both populations. T2D EUR data contained 12171 cases and 56862 controls for a study prevalence of 0.18. T2D EAS data contained 6952 cases and 11865 controls for a study prevalence of 0.37. T2D EUR prevalence was assumed to be 8% while T2D EAS prevalence was assumed to be 9% 53

4.1 Performance of `ursa` with and without haplotype information as compared to `allelecounter` for estimation of reference and alternate allele counts in simulation. ME is the mean error, RMSE is the square root of the mean squared error, and Cor is the correlation of the estimated counts to the simulated counts. . . 60

Acknowledgments

I want to begin by thanking my advisors, Lior Pachter and Noah Zaitlen. Lior has been a wonderful mentor and role model. He taught me what it means for a problem to be "important" and taught me how to critically evaluate science. Noah has been an advocate and close collaborator. He taught me how to present my work to broad audiences without sacrificing rigor, and how to push myself to overcome obstacles I didn't think I could. Most importantly they believed in me and taught me to believe in myself. I want to thank the other members of my committee: Lisa Barcellos and Satish Rao. Lisa invited me to her lab meeting when I first started working on statistical genetics and gave me an opportunity to expand my knowledge of epidemiology. Satish provided an important connection to my roots in theoretical computer science.

It is impossible to provide a complete list of every person that helped me along the way. I have been fortunate to collaborate and discuss science with an enormous amount of people. I'll start by thanking everyone else in the Pachter and Zaitlen labs that I learned from or discussed my projects with: Adam Roberts, Harold Pimmentel, Nick Bray, Aaron Kleinman, Shannon McCurdy, Isaac Joseph, Lorian Schaeffer, Shannon Hateley, Rob Tunney, Bo Li, Danny Park, Shaila Musharoff, Joel Mefford, Meena Subramaniam, and David Siegel. I also spoke extensively with other faculty and students at UC Berkeley, UCSF and elsewhere. Thanks to Ingileif Hallgrmsdttir, Nir Yosef, Jimme Ye, Alkes Price, Nikos Patsopoulos, Bogdan Panaiuc, James Zou, Peter Kraft, Brendan Bulik-Sullivan, Hilary Finucane, Po-Ru Loh, Yakir Reshef, Gleb Kichaev, Huwenbo Shi, Brooke Rhead, Amanda Mok, Gaurav Bhatia and so many more.

Finally, I want to thank my family for supporting me along the way. My mom always encouraged me to follow my dreams and believed that I could do anything. I want to thank my sister, my grandpa and my step-father for flying out to Berkeley to hear my exit talk, knowing full well that they wouldn't understand a word. I want to thank my dog Dorothy for being perfectly happy to sit with me for as long as it took to write this. Most importantly, I want to thank my wife Claire. Claire has been my rock throughout adult life. She was there my first year when I felt like I was in over my head. She's been by my side for every moment, she has helped me pick myself up from defeat, helped me push through when I thought I couldn't, and celebrated my accomplishments with me. I couldn't have done this without her.

Chapter 1

Introduction

1.1 Complex traits and the problem of missing heritability

Complex traits

Human traits, such as disease status and morphological features, can be broadly split into those that do and do not have a genetic component. Among traits that have a genetic component, they can be further subdivided into *mendelian* and *complex* traits. Mendelian traits are controlled by a single genetic variant in a dominant or recessive pattern, where the trait is determined by the presence of a single copy of the disease-causing mutation on one haplotype (dominant) or the presence of the disease-causing mutation on both haplotypes (recessive). Complex traits, on the other hand, are caused by a combination of many hundreds or thousands of genetic variants and the environment. Each individual genetic variant may contribute only a small amount to the trait, where each copy of a relevant mutation that an individual carries increases their height by a small amount, say 1cm, on-average relative to an individual that does not carry the variant. Examples of complex and mendelian traits are given in Table 1.1. As with most topics in biology, this division is not complete, and there are a few traits that are known to be governed by a small number of genes like hair and eye color.

In this manuscript we focus primarily on complex trait *architecture*. Broadly speaking, the architecture of a complex trait is the pattern of genetic variation effecting it. This includes questions like: what proportion of the variance in the phenotype is explained by genetics? what proportion is explained by validated associations? which regions of the genome are enriched for disease-associated variants? how similar are the genetics of two different diseases? how similar are the genetics effects of a disease in two different populations? and related questions.

There are numerous ways to model the relationship between genotype, environment and phenotype. There are also many kinds of genetic variation we can choose to include in our

Table 1.1: Complex traits are caused by hundreds or thousands of genetic variants and the environment, while mendelian traits are effected by a single genetic variant in a dominant or recessive pattern. Complex traits are the focus of this manuscript.

Complex	Mendelian
Height	Sickle-cell disease
Type-II diabetes	Blood type
Rheumatoid arthritis	Lactase persistence
Most cancers	Cleft chin

Table 1.2: Type and number of various kinds of human genetic variation. Single nucleotide polymorphisms (SNPs) are the most common, making up about 95% of all variation. In each case, an example modification to the sequence GATTACA is provided. Note that there are many kinds of structural variation, and the example provided is a copy-number variant.

Type	Example: GATTACA	Number of variants
Single nucleotide polymorphism	GATTGCA	84,387,209
Short insertion or deletion	GACA	3,409,987
Multi-allelic SNP	GATTGCA, GATTCCA	289,480
Structural variation	GAGAGAGAGATTACA	59,797

model, including bi-allelic single nucleotide polymorphisms (SNPs), short indels, multi-allelic SNPs, and various kinds of structural variation (Table 1.2). The majority of human genetic variants are rare, however there are about ten million SNPs that are present in at least 1% of at least one global population [22]. Modeling the impact of *common* SNPs on human phenotypes is the focus of this work, though we acknowledge that evaluating the phenotypic impact of rare genetic variation may be of tremendous medical importance [78, 107].

Let $G \in \{0, 1, 2\}^{N \times M}$ be a matrix of genotypes for N individuals at M SNPs. The number $G_{i,j} \in \{0, 1, 2\}$ represents the number of copies of the minor allele that individual i carries at SNP j. Similarly, let E be an $N \times L$ matrix of L possible environmental effects. Then the most general way to model the relationship between genetics, environment and phenotype is $Y = \Psi(G, E)$ [134]. The function Ψ can include many terms, some of which are listed in Table 1.3. Among these, we will focus on additive genetic variation for reasons that will become clear in the following sections. We will also assume that the trait is quantitative, that is, the the trait is real-valued ($Y \in \mathbb{R}^N$). We will discuss binary (disease) traits in Section 1.2.

Table 1.3: Examples of terms that can be included when modeling the relationship between genotype, environment and phenotype. For the purposes of this manuscript we will focus on additive genetic effects, while acknowledging the potential significance of other kinds of effects in later sections. The ellipsis indicates than in each case we model many additional effects. The notation $\mathbb{1}[C]$ is an indicator variable that the specified condition C holds.

Type	Model
Additive (linear)	$Y = \beta_i G_i + \ldots$
Dominant	$Y = \beta_i \mathbb{1}[G_i > 0] + \ldots$
Recessive	$Y = \beta_i \mathbb{1}[G_i = 2] + \ldots$
SNP-SNP interaction	$Y = \beta_{ij} G_i G_j + \ldots$
SNP-environment interaction	$Y = \beta_{ik} G_i E_k + \ldots$

Missing heritability

The concept of heritability has both an intuitive meaning and a technical definition. In fact, the concept of heritability has several competing technical definitions, which we will be careful to distinguish between in the remainder of this work. At an intuitive level, the concept of heritability relates to the ancient debate between nature and nurture [63]. When we discuss the heritability of a human trait, we think of the relative importance of genetics versus environment in shaping the trait outcome. From a technical standpoint, heritability is the proportion of the variance in the trait that is explained by genetics. In the most general case, the variance of the trait can be partitioned into the genetic variance, the environmental variance, the genetic-environment covariance, and the genetic-environment interaction variance [114]. Specifically, if we partition the trait variance as

$$\sigma_Y^2 = \sigma_G^2 + \sigma_E^2 + 2\sigma_{G,E}^2 + \sigma_{G\times E}^2 \tag{1.1}$$

Then we can define the *broad-sense heritability* as

$$H = \frac{\sigma_G^2}{\sigma_Y^2} \tag{1.2}$$

The genetic component of variance can be further decomposed into the additive, dominant, and epistatic (interaction) components $\sigma_G^2 = \sigma_A^2 + \sigma_D^2 + \sigma_I^2$. This is used to define the *narrow-sense heritability*

$$h_{all}^2 = \frac{\sigma_A^2}{\sigma_Y^2} \tag{1.3}$$

The quantity H^2 represents the total variance in the trait that is explained exclusively by genetics, while the quantity h_{all}^2 represents the total variance that can be explained by additive effects. For the most part, the genetics community is more interested in the narrow-sense

heritability than the broad-sense heritability. There are a number of reasons for this. One is that sharing gene interactions between relatives requires two different genes be identical by descent (IBD). With the important exception of full-sibling and twin relationships, this is relatively rare [114]. Another is that identifying gene interactions is considerably more difficult due to issues of statistical power, a point we will return to in Chapter 2. A third reason is that estimating the broad-sense heritability explained by a set of genetic variants is computationally intractable for arbitrary Ψ [134].

The most widely applied method of estimating heritability is via comparing the correlation of monozygotic (mz) and dizygotic (dz) twins. If we assume that the trait is strictly additive, then we can model the similarity of twins as having a component due to additive genetics (A), common environment (C) and unique environment (E). Since mz twins share 100% of their genome, dz twins share 50% of their genome, and share the same environment, we can estimate heritability by [31]

$$\begin{aligned} r_{mz} &= A + C \\ r_{dz} &= \frac{1}{2}A + C \\ A = h^2_{ACE} &= 2(r_{mz} - r_{dz}) \end{aligned}$$

where r_{mz} and r_{dz} are the phenotypic correlations between mz and dz twins in a population.

Methods that compare the phenotypic resemblance of close relatives to determine heritability can be called *top-down* estimators of heritability. Another approach to estimating heritability is by computing the total variance of the trait explained by a set of discovered variants, the *bottom-up* approach. For a set of uncorrelated genetic variants G', the variance explained is

$$h^2_{G'} = \sum_{i=1}^{|G'|} 2f_i(1 - f_i)\beta_i^2 \tag{1.4}$$

where f_i is the frequency of variant i in the population.

The most common approach to discovering genetic variants to include in the set is the genome-wide association study. In this study, we obtain genotype and trait information for the trait of interest then use linear regression to determine which genetic variants are statistically associated. As previously discussed, there are millions of genetic variants to include in the model. Acquiring the sample size necessary for a multiple regression of millions of regressors to be well-conditioned is problematic, and methods for regularization such as the elastic net[133] are computationally intractable at this scale. Therefore, geneticists resort to computing the association statistic for each SNP in isolation of the remainder of the genome, while conditioning on covariates relevant for the trait of interest. The consequences of this approach will be discussed more thoroughly in Section 1.2.

Each linear regression is a test of the null hypothesis that the SNP is not associated with the trait. To avoid inflating the type-I error rate, the threshold of association must account

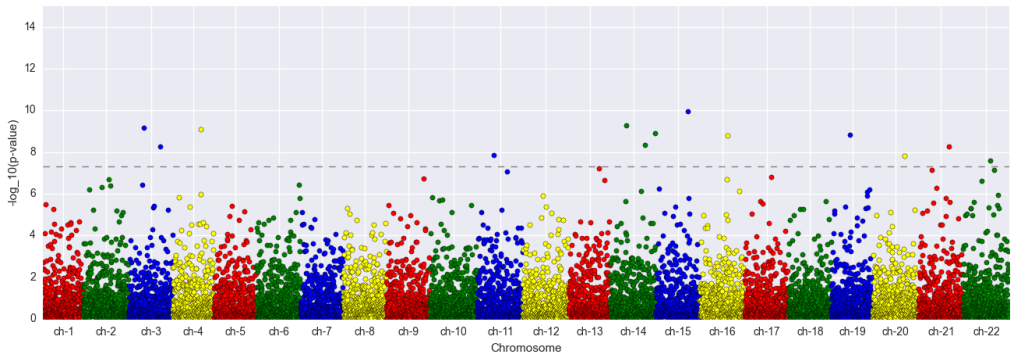

Figure 1.1: An example of a manhattan plot with simulated data. Each of the points is $-\log_{10}(p)$, where the p-value is determined by the χ^2-test statistic of association with the phenotype. The dashed line represents $-\log_{10}(5 \times 10^{-8})$, the threshold of statistical significance in most genome-wide association studies.

for the fact that we are testing millions of correlated hypotheses. Geneticists estimate that there are roughly one-million independent genomic loci in European populations, therefore a Bonferonni multiple-testing correction for a type-I error rate of 0.05 results in a significance threshold of $\frac{0.05}{10^6} = 5 \times 10^{-8}$[89]. The resulting data is usually presented in a Manhattan plot, with highly significant regions showing large log-p-values (Figure 1.1).

We now describe the problem of *missing heritability*. Let G' be the set of un-correlated genetic variants statistically associated in a GWAS for trait Y. Let $h^2_{G'} = h^2_{GWAS}$ be the proportion of phenotypic variance explained by these variants. In it's simplest formulation, the problem of missing heritability is the observation that the variance explained by discovered associations is only a tiny fraction of the heritability estimated by twin studies. That is,

$$h^2_{GWAS} << h^2_{ACE} \tag{1.5}$$

A lot has been written about locating the missing heritability [28, 68, 134, 114, 124]. Notice that we were careful not to call the heritability estimated by twin studies the total narrow-sense heritability, and that we were careful to allow the general model to include gene-gene and gene-environment interactions. Some geneticists have observed that the ACE model commonly used in twin studies implicitly assumes that the trait is strictly additive, and have argued that gene interactions inflate estimates from family studies because close relatives are much more likely to share gene interactions than distant relatives [134]. Others have argued that GWAS are under-powered to detect the many small-effect variants [68, 124]. Furthermore, variants, such as rare SNPs and structural variation, that are not studied in GWAS may contribute to phenotypic variance [68, 28].

1.2 Statistical models for complex trait genetics

Linear mixed models the liability threshold

In the remainder of this work, we will examine statistical models that improve our ability to understand complex trait genetics. The first modeling choice we will make is to only model additive genetic variance, and to assume there are no gene-environment interactions. We also assume that the individuals are only distantly related so that they have no common environment. The next choice we will make is that the SNP effects act via *standardized* genotypes, an assumption we will relax in Chapter 3. That is, let $\mu_G = 2[f_1, \ldots, f_M]$ be the column mean of the genotype matrix G, where f_i is the allele frequency of SNP i. Let $V_G = 2[f_1(1-f_1), \ldots, f_M(1-f_M)]I_M = \text{diag}\left(\frac{1}{M}(G-\mu_G)^\top(G-\mu_G)\right)$ be the allele variances assuming Hardy-Weinberg equilibrium. Now let $X = (G-\mu_G)V_G^{-1}$. Then our model for the complex trait Y is

$$Y = C\gamma + X\beta + \epsilon \tag{1.6}$$

where $\epsilon \perp \beta$ and the $C\gamma$ term allows for covariates to effect the trait mean.

Another assumption we will make for now is that the genetic effects are random, rather than fixed. Specifically, let the vector of genetic effect sizes follow the normal distribution

$$(\beta_1, \ldots, \beta_M) \sim \mathcal{N}\left(0_M, \frac{\sigma_g^2}{M} I_M\right) \tag{1.7}$$

where 0_M is the length-M 0-vector, I_M is the $M \times M$ identity matrix, and σ_G^2 is the trait variance explained by the M variants. This is commonly called the infinitesimal assumption [124]. It has its roots in Fisher's observation that family phenotypic resemblance is consistent with a large number of variants of small effect, and that small effect mutations are more likely to increase fitness[35]. This assumption will be prove very valuable for the remainder of this work. However, some geneticists prefer to assume that the genetic effects are fixed. This approach is especially useful when analyzing short genetic regions where the number of variants makes limiting distribution assumptions problematic [37, 99].

We now have a linear mixed model (LMM). Conditional on fixed effects, the trait is $Y = g + \epsilon$, with g the genetic contribution to the trait ϵ the environmental contribution. By the central limit theorem, g is normally distributed. If we assume that any large environmental effects, such as smoking for lung cancer risk, are known and modeled as fixed-effect covariates, then we can informally argue that the environmental contribution is due to a sum of many small effects and can therefore be assumed normal. This implies that the trait is normally distributed in the population, and indeed this is true for many complex quantitative traits [67]. The distributions of the genetic and environmental contributions are therefore

$$g \sim \mathcal{N}\left(0, K\sigma_g^2\right)$$
$$\epsilon \sim \mathcal{N}\left(0, I_N\sigma_\epsilon^2\right)$$

where $K = XX^\top/M$ is called the *genetic relatedness matrix (GRM)*. The GRM provides an estimate of the shared genetics of distantly related individuals. The trait variance can then be partitioned in order to estimate the total phenotypic variance explained by the set of M SNPs in the genotype matrix

$$h^2_{chip} = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_\epsilon^2} \tag{1.8}$$

which provides a useful lower bound on the narrow sense heritability. However, it does bring many assumptions. In particular, the normalization of the genotype matrix implies that genetic effect sizes are inversely proportional to allele frequency [102]. We will return to this point and relax this assumption in Chapter 3.

We can extend this framework to handle binary (disease) traits by assuming that the trait status is related to an underlying normally-distributed liability via a probit transformation. That is, assume that a disease trait Y that effects $k\%$ of the population has an underlying liability l such that every individual with liability exceed a threshold $t = \Phi^{-1}(1-k)$ has the disease and everyone under the threshold doesn't. This is called the *liability threshold model* [30, 56] (Figure 1.2A).

$$\begin{aligned} l &= G\beta + \epsilon \\ Y &= \mathbb{1}[l > t] \end{aligned}$$

Note that estimates of variance components on the observed (binary) scale must be transformed to get estimates of the variance explained on the underlying liability scale via [26, 56]

$$h^2_{l,chip} = \frac{k(1-k)}{z^2} h^2_{o,chip} \tag{1.9}$$

where $z = \phi(t)$ is the height of the standard normal distribution at the threshold. Furthermore, in most case-control studies of binary traits, the proportion of cases in the study does not match the proportion of cases in the population (Figure 1.2B). In this case, one can show that if the proportion of cases in the sample is p, then the conversion from the observed scale to the liability scale is [56]

$$h^2_{l,chip} = \frac{k^2(1-k)^2}{z^2 p(1-p)} h^2_{o,chip} \tag{1.10}$$

Linear mixed models have broad utility in complex trait genetics beyond providing a lower bound on the amount of trait variance explained by additive genetic variance. Perhaps the most common use of LMMs is to control for population structure in GWAS by explicitly modeling the correlations in the genetic effects that arise from distant family relationships [61]. Another application of mixed models is to find unbiased estimates of the SNP effect sizes that account for the non-random correlation between SNPs in a population, called *linkage disequilibrium* (LD). This is done by first estimating the total genetic (g_i) and environmental (ϵ_i) contribution for and individual and then estimating the effect size of each SNP on the residual variance of the trait[126]. Finally, note the number of variance

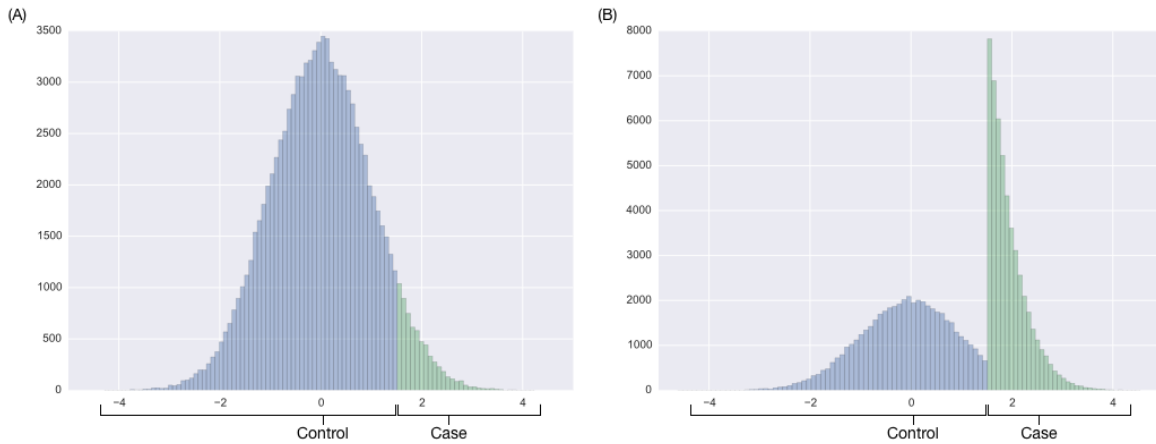

Figure 1.2: Distribution of the underlying liability in a case-control study, using the liability threshold model. In both cases, the population prevalence of the trait is $k = 6\%$ and there are $100,000$ individuals. The threshold at which an individual is considered a case is $\Phi^{-1}(1-k) = 1.55$. (A) With no ascertainment, the underlying liability has a normal distribution. (B) In most studies, there are many more cases than in the general population. In this example, there are $50,000$ cases and $50,000$ controls.

components can be increased to compare the variance explained by different sets of genetic variants

$$
\begin{aligned}
Y &= g_1 + g_2 + \ldots + g_m + \epsilon \\
g_i &\sim \mathcal{N}\left(0, K_i \sigma_i^2\right)
\end{aligned}
$$

where K_i is genetic similarity matrix at subset i of the genetic variants. Gusev et al [39] use this approach to examine how trait variance is partitioned across SNPs in different regulatory regions.

While LMMs have enjoyed remarkable success as a tool for understanding the genetic architecture of complex traits, they are not without their limitations. One limitation is that estimating the kinship matrix has complexity $O(N^2M)$, and is therefore extremely time consuming to estimate for large sample sizes. Similarly, the variance components are usually fit with restricted maximum likelihood estimation (REML) which can be time consuming for large sample sizes and many variance components. That said, a tremendous amount of work has been done to speed up variance component methods [131, 61, 64]

Summary statistics

Another complication to the application of LLMs to complex trait genetics is that the require access to the genotype and phenotype matrices, which are often not provided due to privacy

Table 1.4: Typical summary association data consists of SNP names (rsids), estimates of the effect size and stand error of that SNP, the reference and alternate alleles in the study, and the number of individuals with data at that SNP.

Chr	Pos	rsid	Ref	Alt	N	$\hat{\beta}$	σ_β
1	1199503	rs11260558	T	C	10324	0.0521	0.024
1	1449501	chr1:1449501	A	G	10500	-0.012	0.032
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
22	38896335	rs5757337	A	G	11324	0.102	0.042

concerns [41]. Instead, researched tend to share GWA results data, called *summary statistics*. This usually consists of covariate adjusted effect sizes or odds ratios and their standard errors for a test of association for each SNP against the trait, signed with respect to the reported major allele in the study (Table 1.4). In rare cases, the allele frequencies are also provided. From a computational standpoint, summary statistics can be thought of as a map from genotype-phenotype space to effect size space, which reduces the size of the dataset from $N + M \times N$ to $c \times M$. This dramatically lowers the computational burden of working with the data.

Problematically, summary statistics almost never include correlation matrices of the SNPs in the study. As nearby SNPs are correlated with one another (the aforementioned LD), testing a single SNP implicitly tests all SNPs that are in LD with it. De-correlating estimates of SNP effect sizes requires knowing this correlation structure. This correlation structure can vary dramatically across world populations, but tends to be relatively conserved within each population. This has led some geneticists to develop summary statistics methods that leverage known genotypes from the population of interest to approximate the correlation matrix. These *reference panels* are provided for many populations around the world by the International HapMap Project [38] and the 1000 Genomes Project [22].

We now introduce a formal model for working with summary statistic data by deriving the distribution of the summary statistics. We assume the same linear effect model introduced above. Let Z_i be the Wald test statistic (Z-score) of association for SNP i, $Z_i = \frac{\beta_i}{\sigma_{\beta_i}}$. Then the vector of association statistics Z can be written

$$Z = \frac{X^\top Y}{\sqrt{N}} \tag{1.11}$$

and is asymptotically multivariate normal.

To compute the expected value and variance-covariance matrix of the test statistic, we will assume that the individuals in the study are randomly drawn from population A. That is, assume that in the infinite population-size limit, the SNP correlation matrix is Σ and the

allele frequencies are f_i. Then,

$$\mathbb{E}[X] = 2[f_1, \ldots, f_M]$$
$$\mathbb{E}\left[\frac{X^\top X}{N}\right] = \Sigma_A$$

By the law of total expectation, $\mathbb{E}[X^\top Y] = \mathbb{E}_X[X^\top \mathbb{E}[Y|X]] = 0$. The variance-covariance matrix is

$$\begin{aligned}
\mathbb{E}[ZZ^\top] &= \frac{1}{N}\mathbb{E}[X^\top YY^\top X] \\
&= \frac{1}{N}\mathbb{E}_X[X^\top \mathbb{E}[YY^\top|X]X] \\
&= \frac{\sigma_g^2}{NM}\mathbb{E}[X^\top XX^\top X] + \frac{\sigma_\epsilon^2}{N}\mathbb{E}[X^\top X] \\
&= \frac{\sigma_g^2}{NM}[N(N+1)\Sigma^2 + NM\Sigma] + \sigma_\epsilon^2\Sigma \\
&= \sigma_g^2\frac{N+1}{M}\Sigma^2 + (\sigma_g^2 + \sigma_\epsilon^2)\Sigma
\end{aligned}$$

If we normalize the variance of Y to $\sigma_g^2 + \sigma_\epsilon^2 = 1$, then we have

$$Z \sim \mathcal{N}\left(0, \Sigma + h_{chip}^2\frac{N+1}{M}\Sigma^2\right) \tag{1.12}$$

If we assume we can accurately estimate the population LD matrix Σ, the variance-covariance matrix of Z has only one unknown parameter, the variance explained by the M SNPs h_{chip}^2. This can be exploited to estimate h_{chip}^2. The most common approach for this is called LD-score regression. Let $(\Sigma^2)_{ii} = \sum_{j=1}^M r_{ij}^2 = l_i$ be the *LD score* of SNP i. Then we can estimate h_{chip}^2 via a linear regression of the χ^2 test statistics[7]

$$Z^2 \sim a + bl$$
$$h_{chip}^2 \approx \frac{M}{N+1}b$$

We will extend this model to multiple populations, discuss the consequences of it's assumptions, and consider alternative approaches of fitting the covariance of the distribution of the Z-scores in Chapter 3.

1.3 Gene expression as a genetic trait

By going directly from genotype to phenotype, we are ignoring the complex biological processes that mediate the transition. The field of functional genomics attempts to uncover

how human genetic variation leads to changes in molecular phenotype. This involves the integration of many kinds of genomic data, usually human genotypes or DNA-seq combined with other high throughput *-seq experiments that give information about molecular phenotypes. These include SHAPE-seq [66], which measures RNA structure, ATAC-seq [6], which measures open chromatin regions, CHiP-seq [51, 73], which measures DNA-protein binding, and bisulfite sequencing [72], which measures DNA methylation. There are many more protocols, and new protocols are constantly under development[1].

From among these, the one we will focus on in this manuscript is RNA-Seq, which is a protocol for measuring RNA transcript abundances [76]. RNA-Seq is broadly useful [117], but in the context of this manuscript we will be primarily interested in analyzing how human genetic variation gives rise to variation in transcript and gene abundances. Obtaining accurate estimates of the abundance of genes at the isoform level requires solving numerous computational and statistical challenges. The primary challenges are: read-mapping [108], transcriptome assembly [42], transcript abundance quantification [82] and statistical detection of differential expression between replicates [81]. From among these, we will review transcript abundance quantification and differential expression.

Transcript abundance quantification from RNA-Seq reads

We assume that we have an accurate assembly of the human reference transcriptome, that is, sequences of all isoforms of all genes transcribed in human, and that mapping RNA-Seq reads to the human reference genome or reference transcriptome is accurate, even if it is computationally intensive. With this, quantifying relative transcript abundance from RNA-Seq reads still requires solving many computational and statistical challenges. These are:

- Multi-mapping reads: a read from an exon contained in multiple isoforms of the same gene will map to multiple transcripts in the transcriptome. Futhermore, reads from genes with homology elsewhere in the genome will map to multiple locations in the reference genome.
- Positional bias: fragments are not uniformly sampled from the transcript, which may result from non-uniform fragmentation during library prep.
- Sequence bias: sequences around the beginning and end of transcripts are non-random, meaning that priming and fragmentation strategies result in over-sampling of certain transcripts.

There are numerous methods for estimating gene and transcript level abundances from RNA-Seq reads (see [122, 95, 58, 5, 87, 110, 85] and many others). The most general approach for handling multi-mapping reads while accounting for bias is to use the expectation maximization (EM) algorithm [122, 58, 95, 82, 5]. First, we review the EM algorithm in the general case. Then, we discuss the likelihood-based model generalized in [82] and it's EM algorithm.

Finally, we discuss a recent extension of this approach which dramatically improves speed by observing that the likelihood does not actually require mapped reads [5].

Maximum likelihood estimation and the EM algorithm

Suppose we observe a sequence of n independent and identically distributed (i.i.d) random variables $X_1, \ldots, X_n$ from an unknown distribution $X_i \sim f_0$. In the following, we let x_i denote a random variable and let it's corresponding observation be X_i. Assume that the distribution f_0 belongs to a family of distributions parametrized by θ, $f_0(x) = f(x|\theta)$. Then the distribution from which the observations are drawn can be written

$$f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \tag{1.13}$$

Now by treating the observations as fixed and the parameter as free, we can define the likelihood

$$\mathcal{L}(\theta; X_1, \ldots, X_n) = f(X_i|\theta)^n \tag{1.14}$$

and we can determine the parameter θ which gives the highest probability of observing the data set $X_1, \ldots, X_n$

$$\hat{\theta}_{mle} = \arg\max_{\theta} \mathcal{L}(\theta; X) \tag{1.15}$$

In practice, statistical models are not always completely observed. That is, the likelihood may depend on latent, unknown, variables. For example, if a dataset is drawn from a mixture of two normal distributions, the distribution from which it is drawn is a latent variable. For our application, if a read maps to multiple transcripts, the transcript that generated the read is the latent variable. Formally, let Z be a matrix of latent variables, and let $\mathcal{L}(\theta; X, Z) = f(X, Z|\theta)$ be the likelihood of the complete data. The MLE for the observed dataset is the parameter θ that maximizes the marginal likelihood

$$\mathcal{L}(\theta; X) = \sum_{Z} f(X, Z|\theta) \tag{1.16}$$

However, in practice this quantity can be difficult to compute as the space of possibilities for Z can be exponentially large. The EM algorithm overcomes this by iteratively applying a two-step procedure. In many situations, the distribution of the latent variable can be estimated given a value for the parameter θ and the observations X, while the value of the parameter θ can be estimated easily given complete data Z, X. This is the intuition that is leveraged by the EM algorithm.

We start with an initial guess for the parameter θ_t. Then, in the E-step, we compute the expected value of the hidden variable Z given X, θ_t

$$Z^{(t)} = \mathbb{E}_Z[\mathcal{L}(\theta^{(t)}; X, Z)|X, \theta^{(t)}] \tag{1.17}$$

Then, in the M-step, we compute the MLE estimate of θ given the current guess for the latent variable Z_t

$$\theta^{(t+1)} = \arg\max_{\theta} \mathcal{L}(\theta; X, Z^{(t)}) \tag{1.18}$$

At each step, the resulting likelihood is guaranteed to be at least as large as the previous step [121]. That said, if the likelihood function is not convex, it is only guaranteed to converge to a local maximum.

Likelihood function for RNA-Seq abundance

We now turn to discussing MLE of RNA-Seq abundances. Following [122], assume there are K transcripts each with probability p_k of getting selected, and we have N total reads. If we know which transript each read comes from, this is a multinomial distribution. Specifically, let $Z = \{Z_{i,k}\}_{i=1,k=1}^{N,K}$ be an indicator matrix that read i is from transcript k. Then the likelihood of the probability vector is

$$\mathcal{L}(p; Z) \propto \prod_{k=1}^{K} p_k^{\sum_{i=1}^{N} Z_{i,k}} = \prod_{i=1}^{N} \sum_{k=1}^{K} Z_{i,k} p_k \tag{1.19}$$

However, in practice reads map to multiple transcripts and we don't observe the matrix Z directly. Instead, we observe the matrix $Y = \{Y_{i,k}\}_{i=1,k=1}^{N,K}$ which is an indicator matrix that read i *maps* to transcript k. Then the likelihood becomes

$$\mathcal{L}(p; Y) = \prod_{i=1}^{N} \sum_{k=1}^{K} Y_{i,k} p_k \tag{1.20}$$

and we must infer the true transcript assignments using the EM algorithm. The E-step is

$$Z_{i,k}^{(t)} = \mathbb{E}[Z_{i,k} | Y, p^{(t)}] = \frac{Y_{i,k} p_k^{(t)}}{\sum_{k=1}^{K} Y_{i,k} p_k^{(t)}} \tag{1.21}$$

For the M-step, let $n_k^{(t)} = \sum_{i=1}^{N} Z_{i,k}^{(t)}$. Then

$$p_k^{(t+1)} = \frac{n_k^{(t)}}{N} \tag{1.22}$$

The prior derivation assumes that the frequency of reads from a transcript is directly proportional to it's abundance. However, this is not exactly the case. In the generative mode, first a transcript is selected, then a position on the transcript is selected uniformly at random from which to draw the read. Longer transcripts are more likely to be selected. If the read length is m, then the number of positions in the transcript at which the read can start is $\tilde{l}_k = l_k - m + 1$. The probability of selecting a transcript is therefore

$$\alpha_k = \frac{p_k \tilde{l}_k}{\sum_k p_k \tilde{l}_k} \tag{1.23}$$

Each position in the transcript is chosen uniformly at random, therefore the probability of a specific read is $\frac{\alpha_k}{\tilde{l}_k}$. So that the above likelihood becomes

$$\mathcal{L}(\alpha; Y) \propto \prod_{i=1}^{N} \sum_{k=1}^{K} Y_{i,k} \frac{\alpha_k}{\tilde{l}_k} \tag{1.24}$$

and the abundances can be backed out from the estimated transcript probabilities via $p_k = \frac{\frac{\alpha_k}{\tilde{l}_k}}{\sum_k \frac{\alpha_k}{\tilde{l}_k}}$

More complicated formulations of the likelihood, such as the one used in [94], can model sequence-specific and positional bias while incorporating complex error models. From among these, sequence-specific bias is particularly problematic [95]. In the above formulation of the likelihood, sequence-specific bias can be accounted for by adjusting the effective length. Briefly, one can look at the empirical distribution of 6-mers of the transcript sequence overlapping the 5' fragment and add the bias of each 6-mer on both strands [5, 94]. Notably, the above likelihood cannot account for positional bias and incorporates no error model.

Notice that the above derivation requires that we know which transcripts each read is compatible with. In software like RSEM [58], this is accomplished by mapping each read to the reference transcriptome. This is an extremely time consuming process. Note, however, that the above likelihood doesn't require we know where in the transcript a read came from, just which transcripts it is compatible with. Therefore, if we can determine which transcripts a read maps to without actually mapping the reads, we may be able to speed up abundance estimation dramatically.

Recently, this insight was leveraged by Bray et al [5] in software called `kallisto` using a technique called *k-mer hashing*. Define an *equivalence class* for a read r as the set of transcripts that that read can map to. `kallisto` works via two steps:

- **Index construction:** One can think of the `kallisto` index as a hash table that maps every k-mer present in the transcriptome to the set of transcripts that contain that k-mer. Call this set the k-mer equivalence class.

- **Psuedoalignment:** Each read is shredded into it's corresponding k-mers, and then each k-mer is looked up in the hash table to determine it's k-mer equivalence class. The equivalence class of a read is the intersection of the equivalence classes of the k-mers.

More accurately, the index is a colored transcriptome DeBruijin graph (T-DBG), where each node is a k-mer and each color corresponds to a transcript. Each node is colored by the transcripts that contain that k-mer. The principal difference between the T-DGB and the hash table is that if a sequence of nodes in the T-DBG has the same coloring, those k-mers need not be hashed, and can be skipped.

Bray et al. also show that the likelihood can be re-formulated as a function of the number of reads mapping to each equivalence class.

$$\mathcal{L}(\alpha; Y) \propto \prod_{i=1}^{N} \sum_{k=1}^{K} Y_{i,k} \frac{\alpha_k}{\tilde{l}_k} \tag{1.25}$$

$$= \prod_{e \in E} \left(\sum_{k \in e} \frac{\alpha_k}{\tilde{l}_k} \right)^{c_e} \tag{1.26}$$

where E is the set of all equivalence classes. This formulation of the likelihood also brings a substantial speed improvement. Instead of performing the EM algorithm on millions reads, it is performed on tens of thousands of equivalence classes. This two speed improvements enable non-parametric estimation of the standard error of the abundances via the bootstrap.

Differential expression

If our goal is to understand the molecular path from genotype to phenotype, then the first logical step is to understand how changes in biological condition result in statistically significant changes in transcript abundance. This is the problem of differential expression. Determining what constitutes a statistically significant change in transcript counts is non-trivial because there are two sources of variance, the *biological* variance which arises from the stochasticity of transcription within and between cells [29], and the *technical* variance that arises from stochasticity in quantifying the relative transcript abundances from RNA-Seq.

Thus, an ideal experiment looking for differential expression would contain both biological replicates, where different cDNA libraries constructed from repeated experiments in the same or nearly the same conditions are sequenced, and technical replicates, where the same cDNA library is re-sequenced. In an idealized RNA-Seq experiment, where every read maps uniquely to a single transcript, the observed counts follow a multinomial distribution. This can be well-approximated by a set of independent Poisson random variables, where the variance is equal to the mean, and therefore the variance that would be inferred from technical replicates is redundant [71]. In practice, reads map to multiple transcripts and transcript counts must be inferred via EM as described above. This results in over-dispersion of the count data in technical replicates relative to the Poisson distribution [77]. The most common way to model count data in light of this is via the negative binomial distribution [65, 3].

There are dozens of methods for testing for differential expression (see e.g. [65, 96, 54, 109, 91] and references therein). They incorporate various strategies for shrinkage estimation of the variance parameter in the negative binomial and transformations and normalizations of the count data to fit linear or generalized linear models. A full comparison of the models is beyond the scope of this manuscript, therefore we will focus on the model of [91] which we will further leverage in Chapter 4.

Assume we have measured the transcript abundances of N samples. The logarithm of the counts is approximately normally distributed, therefore let

$$Y_t = X_t\beta_t + \epsilon_t$$

be the log-counts of transcript t as a function of a fixed-effect design matrix of p covariates X, and biological noise $\epsilon_t \sim \mathcal{N}(0, \sigma_t^2 I_N)$. Many RNA-Seq models claim that Y_t is observed [65], but due to the randomness inherent in read alignment, we must also model the technical variance

$$D_t = Y_t + \xi_t = X_t\beta_t + \epsilon_t + \xi_t$$

where $\xi_t \sim \mathcal{N}(0, \tau_t^2 I_N)$ and $\xi_t \perp \epsilon_t, Y_t$. This implies

$$D_t \sim \mathcal{N}\left(X_t\beta_t, (\sigma_t^2 + \tau_t^2)I_N\right)$$

so that accurate estimation of the fixed effects requires accurate estimation of the variance components.

Let c_{ti} be the observed (raw) counts of transcript t in sample i. Then following [3] we normalize the counts using sample specific size factors $\hat{s}_i = \text{median}_t \frac{c_{ti}}{(\prod_{j=1}^N)^{\frac{1}{N}}}$ so that the log-transformed counts are

$$d_{ti} = \log\left(\frac{c_{ti}}{\hat{s}_i} + 0.5\right) \tag{1.27}$$

The technical variance can be estimated from the mean of the sample variances, which are individually estimated via the bootstrap as discussed above, $\hat{\tau}_t = \frac{1}{N}\sum_{i=1}^N \hat{\tau}_{ti}$. Estimating the biological variance is less straightforward. If we let $\hat{\beta}_t = (X_t^\top X_t)^{-1} X_t^\top d_t$ be the OLS estimate of the fixed effects. Then biological variance is the residual variance [91]

$$\hat{\sigma}_t^2 = \max\left(\left(\frac{1}{N-p}\sum_{i=1}^N (d_{ti} - X\hat{\beta})\right) - \hat{\tau}_t, 0\right) \tag{1.28}$$

In many applications, the number of samples N is small and this estimate of the biological variance is unstable. In this situation, most methods employ a shrinkage estimator of the biological variance [91, 3, 54, 109]. However, in this manuscript we are primarily interested in differential expression at the population level and therefore will omit discussion of the shrinkage estimator for variance.

Chapter 2

Local joint testing improves power and identifies hidden heritability in association studies

2.1 Introduction

Genetic association studies typically take a marginal approach to analysis; investigating each SNP in isolation of all other SNPs for association with a phenotype of interest. While this method has led to the discovery of thousands of loci associated with hundreds of phenotypes [119, 27], it fails to capture the additional signal available when multiple SNPs representing independent genetic signals are examined simultaneously [125], or when SNPs are imperfectly imputed [120]. Furthermore, the *hidden* heritability, the difference between the heritability due to genome-wide significant associations and heritability due to genotyped variants, remains substantial [28]. In this work we investigate a local joint testing approach to analysis of genetic data sets in which pairs of variants from the same locus are examined simultaneously for association with a phenotype. The motivation for our approach comes from the mounting evidence that complex traits are highly polygenic [115], that causal variants are not evenly distributed across the genome [40], that known associated loci often harbor multiple causal variants [120, 88, 32, 62, 111, 112, 90], and that the underlying causal variants can be in linkage disequilibrium (LD) with each other [24].

In fact, LD between underlying causal variants can result in additive associations that would be nearly impossible to detect using standard marginal methods. Consider the case of two SNPs: one risk-increasing for a disease, and the other protective. If these SNPs are correlated in the study population then marginal association methods will fail to detect the signal due to the large number of individuals carrying both variants (and therefore having little or no increased risk for the disease). The same effect can occur when the variants have the same effect direction but are anti-correlated in the study population. In the context of this paper, we will refer to these SNP pairs as *linkage masked*. Note that in practice linkage

masking may present in two distinct and important ways. The first is multiple correlated genotyped causal variants with opposite effect direction, which may occur due to the Bulmer effect [9] (see discussion). The second is correlated genotyped variants with opposite tagging of an untyped causal variant. Lappalainen et al [53] give evidence that linkage masking between regulatory and coding variation may be common due to balancing selection. While linkage masked SNPs are difficult to uncover using standard marginal association methods, mixed model heritability is determined by a simultaneous fit of all SNPs while accounting for LD and therefore includes signal from linkage masked SNPs, implicating them as a source of hidden heritability which has not been widely considered [28].

Pairwise (joint) testing for additive effects without a statistical interaction term may help unmask these associations and, more generally, improve power in the presence of gene interactions, multiple causal variants or multiple variants differentially tagging an untyped causal SNP. However, applying joint tests in practice has several problems. Because exact multiple testing correction is usually unknown, several studies have used joint testing for follow up and fine mapping of known associated loci, often revealing additional associated variants [120, 125], and demonstrating the merits of joint testing in practice. Studies such as these are able to ignore multiple hypothesis correction issues due to their focus on known associated regions but do not have the potential to reveal novel loci. Other studies examining genome-wide joint testing of all pairs of SNPs, including those with statistical interaction terms, pay such severe multiple hypothesis correction penalties that many loci found via standard marginal approaches would not reach genome-wide significance [92, 4], and are computationally expensive, though effect methods for reducing the computational burden have been explored [92, 116]. Slavin and Elston [100] proposed testing all adjacent pairs and applied their approach in the WTCCC seven disease study. Howey and Cordell (SnipSnip) [48] proposed using a conditional test on 10 adjacent SNPs to choose a partner SNP for inclusion in the linear model. While these approaches reduce the multiple hypothesis correction penalty we show that they do not capture much of the available power gain. Furthermore, prior approaches have not accounted for a known issue with genotyping error and joint tests [56], which we show impacts these methods.

By testing pairs of SNPs for additive effects, rather than individual SNPs, we improve power to detect loci containing multiple independent causal variants, including those containing linkage masked SNPs. Through local testing, we substantially reduce the multiple hypothesis correction penalty, while simultaneously enriching for situations in which joint tests are more powerful, i.e. when there are independent additive genetic signals contained in each of the SNPs in the test. Rather than employing the overly conservative Bonferroni procedure to account for multiple testing, we extend the work of Han et al. [44] to provide a method for sampling from the null distribution of joint tests orders of magnitude faster than a permutation test, making application computationally efficient. We applied our method to the WTCCC [11] cohorts for bipolar disorder, coronary artery disease, Crohn's diease, hypertension, rheumatoid arthritis, type-1 diabetes and type-2 diabetes. We define *significant locus* as any 1 megabase region containing a statistically significant association at family-wise error rate (FWER) 5% and compare the number of significant loci discovered

from the NHGRI database of replicated associations to the standard marginal method. We compare our approach to SnipSnip [48] and marginal analysis of imputed WTCCC genotypes. We also estimate the phenotypic variance explained by associated SNPs discovered in the marginal and joint approaches. Finally, we apply our method to gene expression data from the gEUVADIS project, comparing the number genes containing *cis*-eQTLs at various false discovery rate (FDR) thresholds under marginal and joint testing approaches.

We find: 1) Local joint testing provides significant power gains when multiple risk variants are proximal, reaching as high as 41% when they are linkage masked. 2) Local joint testing in the original WTCCC cohort discovers seven loci not found via the standard marginal approach, all of which were later replicated in more powerful followup studies. Marginal analysis of imputed genotypes discovers only two of these loci, as well as one locus not detectable in either genotype-based method. 3) New SNPs and loci discovered via local joint testing explain a significant amount of the phenotypic variance of these diseases. 4) joint testing all pairs of *cis*-SNPs in gEUVADIS reveals 607 more genes at FDR 5%, an increase of 10.7% over the marginal approach.

2.2 Methods

We begin by describing the null distributions of the tests in our procedure in order to motivate the sampling approach used to determine the multiple testing correction. We then show that given the marginal Z-scores at two SNPs the joint χ^2 test statistic can be exactly determined. In most cases, determining the significance level of a pairwise test of correlated variables requires a computationally expensive permutation test. However, we build upon a framework for generating marginal test statistics under the null efficiently using a conditional sampling approach [44], which has also been used for genome-wide interaction effect power calculations [116]. This implies that we are able to efficiently sample from the null distribution of the joint test allowing for a dramatic improvement in speed over a permutation test. Finally, we describe the local joint testing procedure. For simplicity, we assume the phenotype and genotypes have been standardized.

Asymptotic distribution of the marginal and joint tests

For many widely used statistical tests, the vector of test statistics over many markers asymptotically follows a multivariate normal distribution (MVN) under the null hypothesis of no association [98, 59]. In particular, let Y be the phenotype of interest and G the genotype matrix, with G_i the genotype at SNP i in a study with N individuals, then the Wald test $Z_i = \frac{\hat{\beta}_i}{s_{\hat{\beta}_i}} = \sqrt{N}\mathrm{Cor}(G_i, Y)$ is asymptotically $\mathcal{N}(0,1)$. From this, one can derive the correlation structure for two tests under the null [44], $\mathrm{Cor}(Z_i, Z_j) = \mathrm{Cor}(G_i, G_j) := \rho_{ij}$, so that the vector of marginal test statistics is asymptotically MVN with mean $\overrightarrow{0}$ and covariance matrix $\Sigma = \{\Sigma_{ij}\}_{i=1,j=1}^{N,N} = \{\rho_{ij}\}_{i=1,j=1}^{N,N}$.

Next, we consider the value of the likelihood ratio test (LRT) statistic for a linear or logistic two-SNP association test. We shoe that it is possible to compute the value of this test statistic directly from the marginal association statistics, without fitting the joint model. Specifically

Observation 1. *Let Z_i, Z_j be the Z-values of test statistics for the SNPs (i, j) against a phenotype Y. Let the correlation between SNPs G_i and G_j be $\rho_{i,j}$. Then the likelihood ratio test statistic for the model $Y \sim 1 + G_i + G_j$ against the null $Y \sim 1$ is*

$$\chi_J^2 = \frac{1}{1-\rho_{i,j}^2}(Z_i^2 + Z_j^2 - 2\rho_{ij}Z_iZ_j) \tag{2.1}$$

and is asymptotically χ_2^2 distributed.

Proof. The equality follows from setting up the normal equations and solving them. Let X be the $N \times 3$ matrix of regressors $X = (\overrightarrow{1}|G_i|G_j)$. Assume that in the linear model $Y = X\beta + \epsilon$, the distribution of the error terms is $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Then the normal equations for the β-coefficients are

$$\beta = (X^\top X)^{-1}X^\top Y$$

solving this and simplifying yields

$$\begin{pmatrix} \beta_1 \\ \beta_i \\ \beta_j \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{N^3}{D}(\rho_{Y,i} - \rho_{ij}\rho_{Y,j}) \\ \frac{N^3}{D}(\rho_{Y,j} - \rho_{ij}\rho_{Y,i}) \end{pmatrix}$$

where

$$D = N^3(1-\rho_{ij}^2)$$

is the determinant of $X^\top X$.

The log-likelihood for a linear model is $\mathcal{L} = -\frac{1}{2\sigma^2}\sum_{i=1}^N (Y_i - \beta^\top X_i)$. Using the above calculation of β, we can find the model log-likelihoods and compute the likelihood ratio

$$\begin{aligned}
\mathcal{L}_N &= -\frac{N}{2\sigma^2} \\
\mathcal{L}_J &= -\frac{N}{2\sigma^2}(1 - 2\rho_{Y,i}\beta_i - 2\rho_{Y,j}\beta_j + \beta_i^2 + \beta_j^2 + \rho_{i,j}\beta_i\beta_j) \\
\chi_J^2 &= 2(\mathcal{L}_J - \mathcal{L}_N)
\end{aligned}$$

Simplifying the above yields,

$$\begin{aligned}\chi_J^2 &= \frac{N}{\sigma^2}\frac{1}{1-\rho_{i,j}}(\rho_{Y,1}^2+\rho_{Y,2}^2-2\rho_{i,j}\rho_{Y,1}\rho_{Y,2})\\ &= \frac{1}{(1-\rho_{i,j})}(Z_{Y,1}^2+Z_{Y,2}^2-2\rho_{i,j}Z_{Y,1}Z_{Y,2})\end{aligned}$$

□

Thus, given the marginal test statistics and the sample correlation of the genotype pair, we can compute the joint test statistic under the null without computationally fitting the model (similar results derived in other contexts can be found in [98] and [125]). This, when combined with MVN sampling of the marginal test statistics, allows a substantial speedup over a permutation test.

While the above is derived in the context of a continuous disease phenotype, it is straightforward to conclude that the framework is extensible to case-control (binary) phenotypes. While we assumed for simplicity the phenotype was standardized, the result is independent of the scale of Y and thus holds for the diseases on the underlying liability scale. Since the least squares model does not rely on the assumption that the error terms are normally distributed (only that they are spherical) the result extends to logistically distributed residuals (logistic regression). In this case the β's are log odds-ratios. We verified computationally by simulating pairs of SNPs at all correlation levels that this equation is exact (not shown).

Estimating the significance threshold and local joint testing

We use a conditional sampling method to sample the marginal test statistics under the null from the multivariate normal distribution. Since distal SNPs are likely to be independent, we choose a window size W_z and 'slide' along the genome, sampling null test statistics at SNP i conditional on the correlation with the previous W_z SNPs [44]. Specifically, the MVN factors as

$$\begin{aligned}f(Z_1,\ldots,Z_L) &= f(Z_1)f(Z_2|Z_1)\ldots f(Z_L|Z_{L-1}\ldots Z_1)\\ &\approx f(Z_1)f(Z_2|Z_1)\ldots f(Z_L|Z_{L-1}\ldots Z_{L-W_z})\end{aligned}$$

and we can sample $Z_i|Z_{i-1}\ldots Z_{i-W_z}$ via the standard conditional MVN

$$Z_i|Z_{i-1}\ldots Z_{i-W_z} = Z_i|Z_p \sim \mathcal{N}\left(\Sigma_{i,p}\Sigma_{p,p}^{-1}Z_p, 1-\Sigma_{i,p}\Sigma_{p,p}^{-1}\Sigma_{i,p}^{\top}\right)$$

where $\Sigma_{i,p}$ is the vector of correlations between SNP i and the conditional SNPs, $\Sigma_{p,p}$ is the $W_z \times W_z$ correlation matrix for the conditional SNPs, and Z_p represents their sampled values.

Each set of sampled marginal null test statistics roughly corresponds to one genome-wide marginal permutation test. Given these marginal null test statistics, we define a joint-testing

window size W_j and compute the joint null test statistics for every pair of SNPs within distance W_j via equation 2.1. These inferred joint null test statistics similarly correspond to a permutation test of all SNP pairs within distance W_j of each other. With these sampled null test statistics in hand, computing the significance threshold is straightforward. For more details, see Algorithm 1, which takes as an input a desired FWER, number of samples (roughly analogous to the number of permutations), a window size, and a set of joint tests T, and outputs a multiple-testing corrected significance level corresponding to the desired FWER. To verify that our method produced results equivalent to those of a permutation test, we performed a permutation test of local joint testing with a window size of 100 SNPs on the first 1000 SNPs of chromosome 1 ($\sim$10 megabases) in the WT control group. We find that the distributions of test statistics under the null for *Jester* and permutation approaches are concordant, and that the significance threshold corresponding to an FWER of 5% are nearly identical (Figure 2.1).

Algorithm 1: Method for sampling from the null distribution to determine significance threshold

Input: significance level α, number of samples n, window size W_z, a set of joint tests T.
Output: Significance threshold
Sample marginal test statistics using a conditional normal approximation
Compute the p-values associated to the marginal tests
for $J_t \in T$ **do**
  Compute $S = \chi^2_{J_t}$ using (2.1)
  Compute the p-value associated to S
end
Sort the p-values from all tests performed
return $(1-\alpha) \times n$th smallest p-value

With the multiple testing correction in hand, the local joint testing procedure is a straightforward modification to the standard GWAS procedure. We choose a window size W_j, correlation cutoff ρ^2_{min} and fit the two-SNP LRT for every pair of SNPs exceeding correlation ρ^2_{min} within W_j markers of each other (Algorithm 2). This procedure is implemented in a

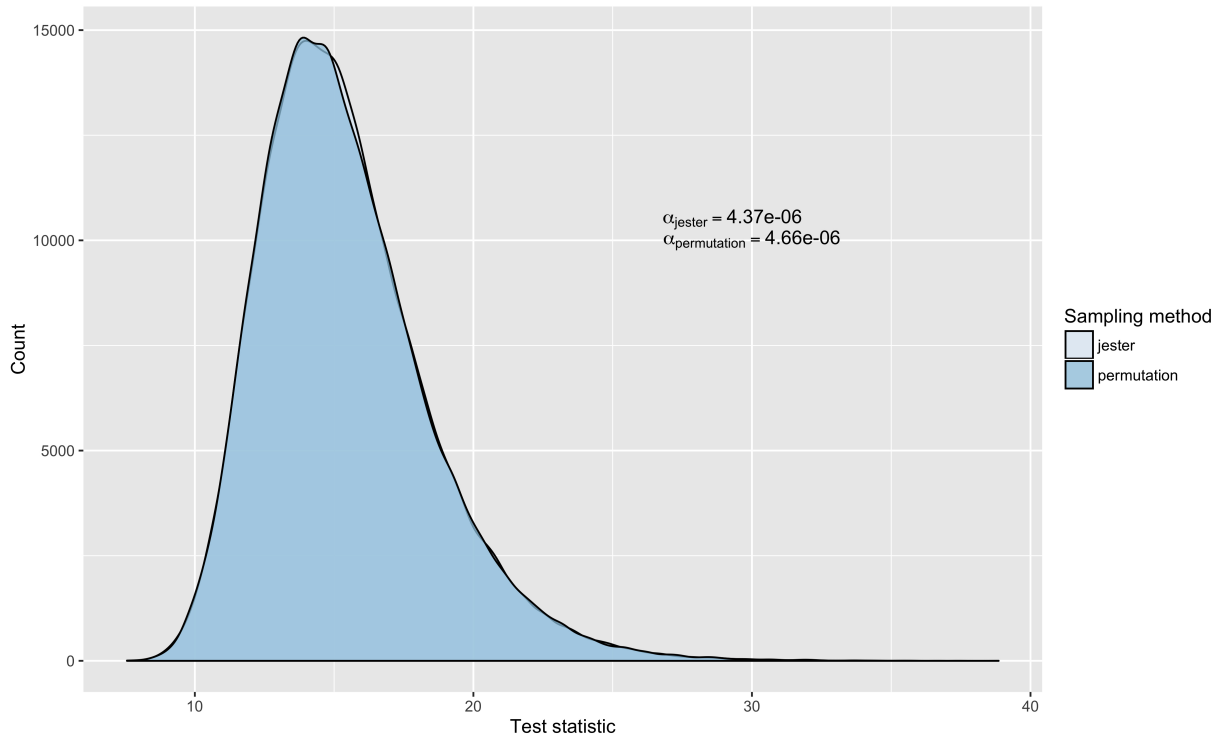

Figure 2.1: Comparison of the density of the maximum test statistic from local joint testing on the first 1000 SNPs of chromosome 1 in the WT controls between *Jester* and permutation approaches. In each case 100K samples were used. The significance level α corresponding to an FWER of 5% for *Jester* in this experiment was $\alpha_{jester} = 4.37\text{e}{-06}$ and for the permutation test was $\alpha_{permutation} = 4.66\text{e}{-06}$. For the purposes of this plot, marginal test statistics (Z values) and joint test statistics (χ^2_2) were transformed to χ^2_1.

python package called *Jester* (`github.com/brielin/Jester`).

Algorithm 2: *Jester*'s GWAS pipeline

Input: A matrix of genotypes G and a vector of phenotypes Y for N individuals
Output: A set of pairs of variants meeting genome-wide significance
Perform standard QC on G and Y
Estimate the joint test significance threshold α_{LJT}
for *SNP* $G_i \in G$ **do**
  Test G_i for association with Y
  Jointly test G_i with any of the preceding W_j markers correlated with G_i at level ρ^2_{min} or greater
end
return SNP pairs with p-value $< \alpha_{LJT}$

Filtering false positives

Lee et al. [56] describe an issue where genotyping errors that would go unnoticed in standard QC procedures can cause inflation in joint and conditional tests of association. When SNPs are highly correlated, mis-called bases in only the cases or controls induce rare haplotypes. As these haplotypes are only present in the cases or controls, this increases the association signal in the joint test [56].

While performing our analysis, we found many highly correlated ($|\rho| > 0.9$) pairs of SNPs where neither SNP had substantial marginal signal but together showed an extremely strong association. We accounted for this in two ways: first, we considered only associations arising from pairs with correlation less than 0.9, second, we used imputation against 1000 genomes to reanalyze jointly significant genotyped SNPs. Specifically, for each pair of potentially significant SNPs, we used the `-pgs` flag of *Impute2* to hold out the genotyped SNPs and replace them with values imputed from the surrounding SNPs. We then re-computed the test statistic using the imputed values of the SNPs. When the signal was a true association, the joint test statistic remained significant after pgs imputation (Table 2.4). However when the association appeared to be driven by genotyping error, the joint test statistic became insignificant after pgs imputation (Table 2.7). In this way, we overcome the false positive error identified by Lee et al. [56].

Datasets

We analyzed the WTCCC phenotypes bipolar disorder, Crohns disease, coronary artery disease, hypertension, rheumatoid arthritis, type-1 diabetes and type-2 diabetes (CD, CAD, HT, RA, T1D, T2D). We chose this data set because it was one of the first GWAS performed and the phenotypes have been subsequently studied in independent large scale GWAS. Thus, we emphasize the potential of early discovery of true effect leveraging non-standard GWAS methods. We used a window size of $W_z = W_j = 100$ SNPs for estimating the null distribution, and a correlation cutoff of $\rho^2_{min} = 0$ (all pairs in the window). We performed standard QC on the data, removing individuals with missingness > 0.1, SNPs with missing-ness > 0.1, markers failing a Hardy-Heinberg equilibrium (HWE) test at significance level 0.001, and SNPs with minor allele frequency < 0.05. To impute the WTCCC cohort, genotypes were split into 1 mega-base regions and pre-phased against the 1000 Genomes EUR reference panel using HAPI-UR, then imputed using Impute2 against the same reference panel. Non-biallelic SNPs and SNPs with reference panel frequency below 5% were not imputed. All imputed SNPs with info score below 0.5 were excluded from further analysis.

We also analyzed gene espression data for 16155 genes of the the gEUVADIS European dataset. Raw RNA-sequencing reads obtained from the European Nucleotide Archive were aligned to the transcriptome using UCSC annotations matching hg19 coordinates. RSEM was used to estimate the abundances of each annotated isoform and total gene abundance is calculated as the sum of all isoform abundances normalized to one million total counts or transcripts per million (TPM). Genotyping data was obtained from the 1000 Genomes

Phase III public release. eQTL mapping was performed on a per-gene basis. The *cis* region of each eQTL was defined as all SNPs with MAF $> 5\%$ within 200KB of the transcription start site (TSS), which was chosen because the vast majority of eQTLs are known to be contained in this region [104]. Joint tests were performed between all pairs of SNPs in the *cis* region. In each analysis, 30 genotype principal components were included as covariates. Approximate permutation tests from our sampling procedure with 2500 samples were used to infer permuted p-values separately for the marginal and joint approaches, which were then independently analyzed to determine the number of significant genes at FDR 1%-25%.

2.3 Results

Multiple Testing Penalties for WTCCC and Power

We computed the multiple testing correction for local joint testing in the WTCCC cohort using *Jester* for various window sizes and ρ^2_{min} cutoffs. We define the *effective number of tests (ENT)* as the number of independent tests that would correspond to a corrected significance level of α_C for corresponding FWER α. That is, $ENT = \alpha/\alpha_C$. We found that the significance level of genome-wide marginal testing at FWER 5% was $\alpha_M = 2.3 \times 10^{-7}$. We chose to use a window size of 100 and ρ^2_{min} cutoff of 0 for our main analysis, increasing the ENT by a factor of 19.55 over the marginal test, even though we perform 100 times as many tests ($\alpha_J = 1.18 \times 10^{-8}$). We chose the window size based on the work of Han et al. [44], and used no correlation cutoff because our power simulations showed an increase in power even in the absence of LD (Figure 2.2, right). For smaller window sizes and larger cutoffs, the multiple testing burden decreases substantially. Using a modest ρ^2_{min} cutoff of 0.004, for example, increases the number of tests by a factor of 8.90 over the marginal test ($\alpha_{J,0.004} = 2.5 \times 10^{-8}$).

We sought to determine the relative power of local joint testing versus marginal testing in the case of 1) two correlated variants in LD and 2) a single causal variant. We simulated pairs of genotypes for 5000 individuals with allele frequency 50% and correlation $\rho \in [-1, 1]$. We simulated phenotypes in a linear model with standard normally distributed environmental noise where 1) both SNPs had an effect size of $\beta = 0.1$ or 2) only SNP 1 had an effect size of $\beta = 0.1$. We used the significance thresholds computed above to incorporate the effect of testing pairs of correlated variants genome-wide. We found a substantial increase in power for modest window sizes and small correlation cutoffs (Figure 2.2, right) when there were multiple causal variants. Using a window size 50 and cutoff of 0.004 lead to an increase in power of up to 41.0% when there were multiple correlated causal variants, while a window size of 100 and cutoff of 0.0 saw an increase in power of 35.4%. In the absence of multiple causal variants, the increase in multiple testing burden and degree-of-freedom penalty gave a decrease in power of 15% and 20%, respectively for the two testing contitions, for all correlation levels. Note that while our method shows its most substantial gain when SNPs are linkage masked, we also see up to a 25% increase in power when the causal variants are

uncorrelated.

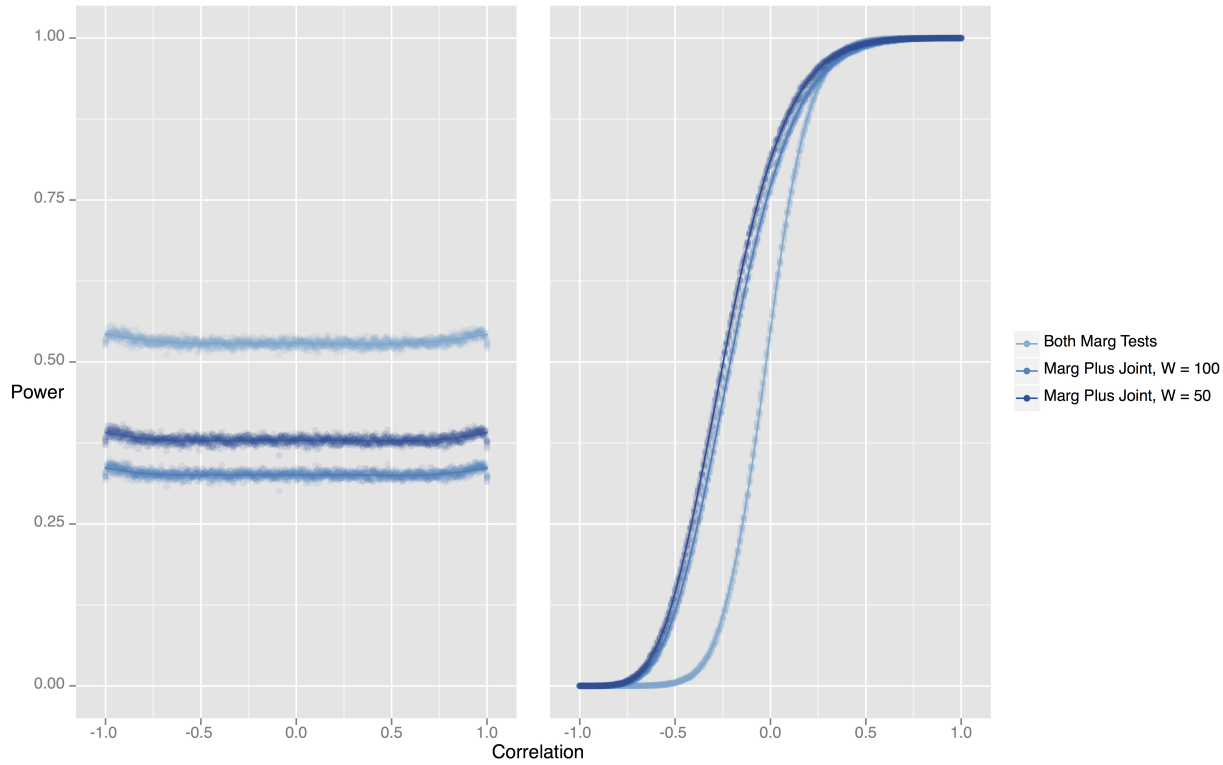

Figure 2.2: (Left) Joint testing genome-wide shows a power loss for all correlation structures when only one SNP affects the trait. (Right) Joint testing genome-wide shows a substantial power gain for anti-correlated SNPs that both affect the trait.

Significant loci in WTCCC

We analyzed the WTCCC phenotypes using four different methods: 1) marginal tests at an FWER of 5% (2.3×10^{-7}), 2) joint tests with window size 100 and $\rho^2_{min} = 0$ at the significance level estimated by *Jester* corresponding to an FWER of 5% (1.18×10^{-8}), 3) SnipSnip [48] with window size of 10 SNPs at significance level 5×10^{-8} and 4) marginal tests of imputed genotypes at significance level 5×10^{-8}. Since SnipSnip does not include an analysis of the multiple testing correction we chose to use their suggested significance threshold.

Local joint testing resulted in the discovery of 2.3 times as many associated SNPs over the marginal method, summarized in Table 2.1. In Figure 2.3 we provide a plot of the density of the correlation between the pair of SNPs that are genome-wide significant at FWER 5% using *Jester*. These SNP pairs have a range of correlations, but SNP pairs where neither SNP was discovered in marginal testing have higher correlation (signed with respect

to opposite effect directions) than the average (Figure 2.3). The marginal test of genotyped SNPs revealed 17 significant loci. *Jester* discovered seven additional loci while missing two of the original due to the increased multiple testing burden, for a total of 22 significant loci. For each of these seven newly significant loci, we searched the NHGRI GWAS database [119] for reported associations and found that each had been reported in more powerful disease-specific follow-up studies. The significant SNP pairs in four of these seven novel loci were linkage masked, with correlations ranging from 0.26 to 0.74 (signed w.r.t opposite direction SNP effects). Interestingly, SnipSnip discovered fewer SNPs and loci than the marginal method, but we emphasize that the set of loci it uncovered are not a strict subset of those discovered via the marginal method (Table 2.5). That is, it discovers some new loci while missing some that are marginally significant and thus remains useful as a secondary analysis tool.

We also compared our method to a GWAS of imputed genotypes, the current gold standard method. This allows us both to determine how our method compares in the number of discovered loci, and whether the linkage masked loci that *Jester* uncovered were due to correlated SNPs with opposite tagging of an untyped causal variant. The imputed GWAS discovered 20 loci in total: the 17 marginally discovered loci, two of the seven linkage masked loci (Table 2.2) , and one locus found by SnipSnip but not *Jester* or the marginal GWAS (CD 18p11.21, Tables **??**). The two loci discovered by both *Jester* and imputation (CD 5q31.1 and T1D 10p15.1) were linkage masked, but the discovery of a significant SNP after imputation implies this was due to opposite tagging of an untyped causal. Of the remaining five loci, three (CD 10q21.3, RA 1p36.32, RA 10p15.1) were uncorrelated, supporting the presence of multiple causal variants. The final two loci (CD 6p21.32, T2D 9p21.3) were correlated ($r = 0.26$ and $r = 0.51$, respectively) but did not contain a significant SNP after imputation. We conclude that these loci are strong candidates for followup study to validate the presence of linkage masking. For complete details of all associations discovered in each method, see Tables 2.3-2.6.

Phenotypic variance explained by new genome-wide significant associations

We sought to quantify the effect of newly significant SNPs on the phenotypic variance explained by to genome-wide significant associations. We used *GCTA* to compute the genetic relationship matrix (GRM) for each phenotype using only SNPs identified as significant using either the classic marginal or local joint testing method. We used the *GCTA* `-mgrm` mode to fit a mixed model containing both marginal and joint GRM's, and performed a likelihood ratio test of the full model against a reduced model containing only the marginal GRM to assess statistical significance. We were unable to fit the model for RA because of numerical issues. For the four remaining disease phenotypes with genome-wide significant associations, three (CD, T1D, T2D) show a statistically significant increase in the phenotypic variance explained, while one (CAD) shows no increase in the phenotypic variance explained. The

Table 2.1: (Left) Total number of loci containing genome-wide significant SNPs discovered using standard marginal, local joint, SnipSnip (SS), and genome-wide imputation (imp) testing methods. (Right) Total number of genome-wide significant SNPs discovered using standard marginal, local joint, SnipSnip (SS) and genome-wide imputation (imp) testing methods. For our analysis of T1D and RA, we removed chromosome 6 because of the large effect HLA locus.

	marg	Jester	SS	imp	marg	Jester	SS	imp
BD	0	0	0	0	0	0	0	0
CAD	1	1	0	1	16	24	0	96
CD	7	8	5	8	58	89	56	587
HT	0	0	0	0	0	0	0	0
RA	2	4	3	3	6	29	6	88
T1D	5	6	3	6	14	82	14	264
T2D	2	3	2	2	16	25	24	76
Total	17	22	13	20	110	249	100	1111

Table 2.2: Loci that were not significant in the standard marginal approach but became significant using *Jester*. ρ indicates correlation of SNPs signed with respect to opposite effect direction. Results at these loci from imputation against 1000 genomes are also reported. P-values which are significant for a particular testing method are denoted by an asterisk.

		marginal		Jester				imputation	
Dis	Locus	SNP	PV	SNP1	SNP2	ρ	PV	SNP	PV
CD	5q31.1	rs6596075	5.97E-07	rs6596075	rs273913	0.32	9.34E-09*	rs6897597	3.22E-8*
	6p21.32	rs9469220	9.09E-07	rs9469220	rs12524063	0.26	2.81E-10*	rs210194	3.359E-6
	10q21.3	rs10761659	2.82E-07	rs10995271	rs10995271	0.01	3.02E-09*	rs10761659	1.88e-07
RA	1p36.32	rs10910099	3.09E-06	rs12027041	rs10910099	0.00	1.02E-09*	rs867436	4.795e-07
	10p15.1	rs2104286	7.31E-06	rs1570527	rs2104286	0.03	3.71E-09*	rs2181623	5.182e-05
T1D	10p15.1	rs2104286	8.28E-06	rs12722489	rs2104286	0.74	1.17E-08*	rs12722563	6.39E-09*
T2D	9p21.3	rs523096	2.49E-04	rs10757283	rs10811661	0.51	3.36E-09*	rs12555274	1.877e-07

Table 2.3: Loci containing a marginally significant SNP at the 0.05 level after correction for genome-wide multiple testing (p ¡ 2.18e-7)

Disease	Locus	RSID	Pos	Beta	Beta SE	p-Value
CAD	9p21.3	rs1333049	22115503	0.636	0.0834	2.31E-14
CD	1p31.3	rs11805303	67387537	-0.634	0.0886	8.09E-13
	2q37.1	rs10210302	233940839	0.656	0.088	9.06E-14
	5p13.1	rs17234657	40437266	-0.859	0.118	3.35E-13
	5q33.1	rs11747270	150239060	-0.839	0.154	5.31E-08
	10q24.2	rs10883365	101277754	-0.49	0.0865	1.52E-08
	16q12.1	rs2076756	49314382	-0.731	0.0943	8.78E-15
	18p11.21	rs2542151	12769947	-0.594	0.109	5.13E-08
RA	1p13.2	rs6679677	114015850	-1.34	0.129	2.92E-25
	MHC	rs6457617	32771829	1.64	0.0909	2.10E-72
T1D	1p13.2	rs6679677	114015850	-1.31	0.124	8.24E-26
	MHC	rs9268877	32539125	1.82	0.101	2.29E-73
	12q24.13	rs17696736	110949538	-0.663	0.084	2.90E-15
	12q13.2	rs11171739	54756892	-0.562	0.0833	1.58E-11
	16p13.13	rs12924729	11095284	0.502	0.09	2.40E-08
T2D	10q25.2	rs4506565	114746031	-0.63	0.0877	7.12E-13
	16q12.2	rs7193144	52368187	-0.482	0.0852	1.56E-08

increase in phenotypic variance explained varies by phenotype, with CD showing a 74% increase, T1D showing a 12% increase, and T2D showing a 75% increase (Figure 2.3).

However, winners curse may disproportionally impact joint testing due to the increased number of tests relative to the marginal approach. To quantify the effect of winners curse we estimated the out of sample phenotypic variance explained by genome-wide significant associations for T2D using European individuals from the Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort (detailed description of the cohort and study design can be found in dbGaP, Study Accession: phs000674.v1.p1). Of the 16 marginally significant SNPs present in WTCCC, 5 were also present in GERA. Of the 250 joint significant SNPs present in WTCCC, 65 were also present in GERA. In the GERA analysis, we estimated the phenotypic variance explained by those sets of SNPs as the coefficient of determination (R^2) in a linear regression of the marginal and joint SNP sets against the T2D phenotype. We found that the variance in liability explained in GERA for the marginally significant SNPs was 1.03% (0.07%), and the variance in liability explained in GERA for the joint significant SNPs was 1.52% (0.11%), an increase of 46.6% ($p = 0.0049$ for an LRT of the joint and marginal SNPs against just the maginal SNPs). Therefore we conclude that winners curse does effect in-sample estimates, but the increase in variance explained due to linkage masked variants remains substantial.

Table 2.4: Loci significant in the WTCCC consortium at level 0.05 after correction for multiple testing of all SNPs with their 100 closest neighbors. A reference to the NHGRI study first reporting the association is provided.

Disease	Locus	Pos1	Pos2	AF1	AF2	Corr	Marg pv 1	Marg pv 2	Joint P	Imp corr (r)	Imp pv 1	Imp pv 2	Imp pv joint	NHGRI study first reported
CAD	9p21.3	22124000	22116000	0.2075	0.247	-0.0768	0.000343	2.31E-14	1.67E-15	0.0831	0.00094	5.24E-14	8.88E-15	WTCCC
CD	1p31.3	67371000	67368000	0.062	0.2235	0.417	0.0454	7.53E-06	3.96E-09	0.422	0.0831	8.58E-06	1.39E-08	WTCCC
	2q37.1	233940000	233720000	0.2255	0.057	0.0162	9.06E-14	0.000797	1.33E-15	-0.0191	8.09E-14	0.00202	2.55E-15	WTCCC
	5p13.1	40610000	40379000	0.051	0.1095	-0.117	0.000353	8.94E-07	5.99E-10	-0.115	0.000503	2.96E-07	2.75E-10	WTCCC
	5q31.1	131770000	131690000	0.0755	0.183	-0.322	5.97E-07	0.115	9.34E-09	-0.287	4.82E-07	0.096	1.04E-08	Barret 2008 NG
	6p21.32 (MHC)	32766000	32405000	0.2495	0.043	-0.259	9.09E-07	0.00272	2.81E-10	-9	-9.00E+00	-9.00E+00	-9.00E+00	Yamakazi 2012 GI
	10q21.3	64192000	64108000	0.032	0.2035	0.0108	0.000344	4.75E-07	3.02E-09	0.0141	0.000567	3.04E-06	2.91E-08	Rioux 2007 NG
	10q24.2	101310000	101280000	0.1255	0.117	0.00789	3.19E-07	0.00105	7.99E-09	0.00912	8.16E-07	0.00108	2.13E-08	WTCCC
	16q12.1	49317000	49309000	0.17	0.188	-0.534	0.126	2.78E-07	5.50E-12	0.539	0.12	6.59E-07	1.41E-11	WTCCC
RA	1p13.2	113840000	113600000	0.123	0.175	0.0837	4.10E-07	3.06E-05	5.82E-11	-0.0889	1.78E-07	2.25E-05	1.51E-11	WTCCC
	1p36.32	3614600	2565700	0.2155	0.164	-0.000202	1.11E-05	3.09E-06	1.02E-09	0.0312	0.153	3.10E-06	4.85E-06	Raychaudhuri 2008 NG
	6p21.32 (MHC)	32885000	32850000	0.1985	0.154	0.799	0.00266	1.82E-05	0	-9	-9.00E+00	-9.00E+00	-9.00E+00	WTCCC
	10p15.1	6444700	6139100	0.088	0.135	-0.026	3.60E-05	7.31E-06	3.71E-09	0.0317	4.32E-05	3.52E-06	1.90E-09	Stahl 2010 NG
T1D	1p13.2	113630000	113600000	0.08	0.0505	0.753	3.25E-06	0.799	2.31E-12	0.769	1.91E-05	0.799	3.43E-11	WTCCC
	6p21.32 (MHC)	113630000	113600000	0.162	0.1275	0.661	5.52E-07	0.0171	0	-9	-9.00E+00	-9.00E+00	-9.00E+00	WTCCC
	10p15.1	6142000	6139100	0.0845	0.135	0.74	0.519	8.28E-06	1.17E-08	0.758	0.398	1.15E-05	2.92E-08	Barret 2008 NG
	12q13.2	56109000	54757000	0.12	0.2255	-0.0425	0.00558	1.58E-11	4.87E-12					WTCCC
	12q24.13	110910000	110700000	0.063	0.153	-0.242	0.00271	3.58E-07	1.41E-10	0.262	0.000401	1.01E-07	1.00E-12	WTCCC
	16p13.13	11311000	11101000	0.091	0.2035	-0.118	0.000289	6.79E-06	5.21E-09	-0.118	0.000279	1.23E-05	9.48E-09	WTCCC
T2D	9p21.3	22124000	22124000	0.206	0.078	0.505	0.00445	0.000769	3.36E-09	0.558	0.022	0.000973	1.81E-08	Scott 2007 Science
	10q25.2	114800000	114750000	0.075	0.176	-0.144	0.14	7.12E-13	2.12E-13	-0.154	0.214	9.71E-13	4.28E-13	WTCCC
	16q12.2	52358000	52071000	0.2245	0.231	-0.00486	2.96E-07	0.000244	2.30E-09	-0.00463	2.71E-07	0.000815	6.50E-09	WTCCC

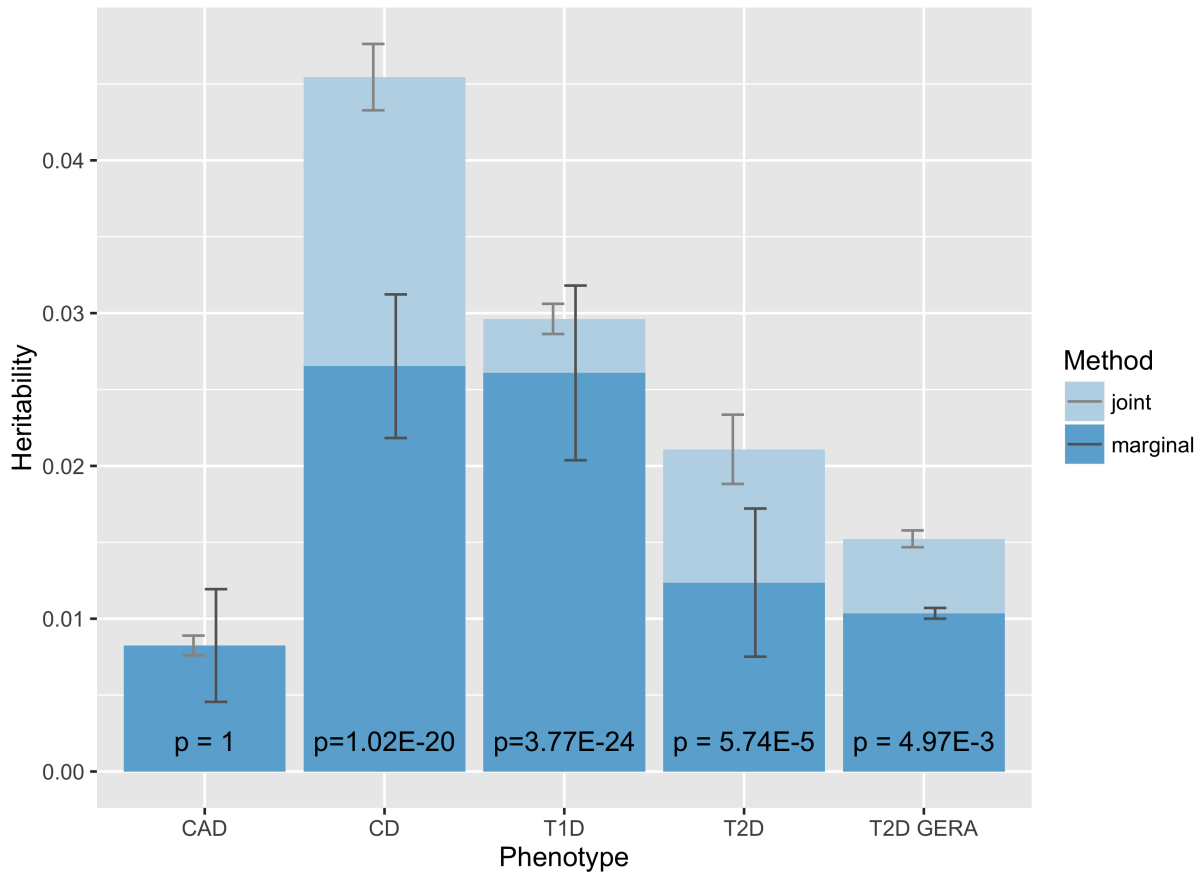

Figure 2.3: The heritability of liability due to genome-wide significant marginal associations (dark blue) plus additional heritability explained by genome-wide significant joint associations (light blue). Error bars correspond to the standard error of the heritability estimates. In all cases but the T2D GERA, p-values correspond to the liklihood ratio test of the linear mixed-model fit with both marginal and joint GRM against the linear-mixed model fit only with the marginal GRM. In the T2D GERA case, the p-value corresponds to a likelihood ratio test of the linear model fit will joint and marginal significant SNPs against the model fit with only marginally-significant SNPs.

Table 2.5: Results of running SnipSnip on the WT disease cohort using the default parameters. All pairs with correlation above 0.8 were removed to filter false positives.

Disease	Locus	Position 1	Position 2	Corr	χ^2-M	PV Marg	χ^2-J	PV Joint
CD	1	67406223	67416128	0.32	47.35	5.93E-12	37.43	9.44E-10
	2	233943769	234015235	0.29	56.03	7.11E-14	54.28	1.73E-13
	5	40437266	40438290	0.24	52.87	3.56E-13	76.27	2.46E-18
	5	131829057	131815177	0.13	23.12	1.52E-06	32.54	1.16E-08
	6	32766288	32771829	0.32	24.25	8.45E-07	27.63	1.47E-07
	16	49323628	49317048	0.29	32.85	9.91E-09	58.84	1.71E-14
	18	12824359	12802167	0.33	9.65	1.88E-3	27.48	1.58E-07
RA	1	114015850	113930493	0.39	109.27	1.41E-25	84.57	3.71E-20
	6	32471505	32467409	0.14	273.18	2.29E-61	287.63	1.63E-64
	11	3638426	3635022	0.47	15.41	8.65E-05	37.39	9.67E-10
T1D	1	114015850	113930493	0.40	112.54	2.72E-26	73.70	9.07E-18
	6	31726100	31730585	0.32	324.81	1.29E-72	560.78	5.66E-124
	12	54756892	54841289	0.15	45.79	1.31E-11	42.85	5.90E-11
T2D	9	22124094	22124172	0.25	11.46	7.1E-4	29.25	6.35E-08
	10	114779067	114795850	0.18	48.94	2.63E-12	49.28	2.22E-12
	16	52368187	52365759	0.39	32.12	1.45E-08	33.18	8.37E-09

gEUVADIS eQTL analysis

We compared the number of genes containing an eQTL at an FDR of 1%-25% using standard marginal linear regression of all *cis*-SNPs against joint tests of all *cis*-SNPs (Table 2.8). At an FDR of 5%, we find that 5641 of 16155 genes contain an eQTL using the marginal test, and 6248 genes contain an eQTL using the joint test, an increase of 10.7%. As in our analysis of the Wellcome Trust data, the genes discovered using the marginal approach are not a strict subset of those discovered using the joint approach. For each level of FDR, we determined the proportion of genes uncovered in the joint but not marginal approach that appear to be linkage masked. At FDR 5%, the joint testing approach discovers 908 new genes. In 381 of those 908 genes the significant SNP pair have a correlation of greater than 0.2 signed with respect to opposite direction SNP effects (Table 2.8). In Figure 2.3 we provide a plot of the density of the correlation between the pair of SNPs in the most significant pair for each gene containing an eQTL at FDR 5% using *Jester* These SNP pairs have a range of correlations, but SNP pairs in genes without an eQTL discovered in marginal testing have higher correlation (signed with respect to opposite effect directions) than the average (Figure 2.3).

Table 2.6: Replicated loci found in a GWAS on WTCCC variants imputed to 1000 genomes. The Comment column indicates whether the locus was detected in the marginal or 100-SNP joint method, and provides a reference to the earliest replecation for the loci not found in the standard marginal or local joint approaches

DIS	LOC	BP	Imp AF	χ^2	pv	Comment
CAD	9p21.3	22018781	0.4297	29.97	4.39E-08	Marg
CD	1p31.3	67552639	0.06555	30.57	3.22E-08	Marg
	2q37.1	234143048	0.4826	47.3	6.08E-12	Marg
	5p13.1	40319877	0.1225	48.53	3.25E-12	Marg
	5q31.1	131743465	0.2415	30.57	3.22E-08	Joint
	5q33.1	150240076	0.06646	30.42	3.49E-08	Marg
	10q24.2	101277816	0.4685	30.24	3.82E-08	Marg
	16q12.1	50737498	0.2751	31.25	2.27E-08	Marg
	18p11.21	12774326	0.1635	31.11	2.44E-08	Parkes et al 2007 NG
RA	1p13.2	114173410	0.1201	50.28	1.34E-12	Marg
	MHC	29690056	0.2316	39.1	4.02E-10	Marg
	6q23.3	138173422	0.07418	30.09	4.11E-08	Okada et al 2014 Nature
T1D	1p13.2	114075796	0.2287	30.8	2.86E-08	Marg
	MHC	32096001	0.1256	251.6	1.17E-56	Marg
	10p15.1	6069561	0.1181	33.71	6.39E-09	Joint
	12q13.2	56379060	0.416	30.07	4.17E-08	Marg
	12q24.13	112486818	0.4243	62.73	2.37E-15	Marg
	16p13.3	11164567	0.3222	31.46	2.04E-08	Marg
T2D	10q25.2	114732906	0.4478	29.97	4.39E-08	Marg
	16q12.2	53800954	0.4074	30.21	3.87E-08	Marg

2.4 Discussion

In this work we described a local joint testing procedure, its multiple testing correction, the genetic architecture for which it is well powered, a system for reducing susceptibility to genotyping error, and implications for phenotypic variance explained from GWAS SNPs. We have shown that when loci harbor multiple causal variants, the joint test can outperform the marginal test substantially. We observed that our method out-performs the standard marginal association method, lending further evidence that disease loci frequently harbor multiple causal mutations. In our simulations, we find that the largest power gains come from linkage masked SNPs: when SNPs have opposite effect direction but are correlated in the study population. Furthermore, four of the seven (57%) newly significant WT loci have the aforementioned property and 381 of the 908 (42%) newly discovered eQTLs at FDR 5% show evidence of linkage masking. This finding is weakened only somewhat by the fact that two of the four linkage masked loci in WTCCC appear to be type 2 linkage

Table 2.7: A sample of loci which appeared to be false positives in our dataset. In all cases, marginal signal was eliminated after replacing genotypes with those estimated from imputation against 1000 genomes.

Disease	Locus	Pos1	Pos2	AF1	AF2	Cor	Marg pv1	Imp pv1	Info score	Marg pv2	Imp pv2	Info score
BD	1p35.2	31848000	31838000	0.0555	0.068	0.837	0.00273	0.684	0.955	0.49	0.571	0.945
	6q21	107090000	107090000	0.1935	0.168	0.82	0.816	0.784	0.973	9.66E-05	0.284	0.98
	10p11.1	38113000	37979000	0.044	0.0385	0.85	0.145	0.916	0.917	0.0117	0.363	0.931
	21q22.2	39348000	39342000	0.029	0.027	0.834	0.00067	0.326	0.845	0.401	0.384	0.903
CAD	1p13.1	108370000	108360000	0.027	0.0345	0.84	0.841	0.796	0.948	0.00151	0.72	0.81
	2q14.2	121520000	121510000	0.178	0.141	0.802	7.91E-06	0.00537	0.953	0.266	0.0449	0.955
	3q25.33	161940000	161890000	0.164	0.1325	0.824	0.0446	0.466	0.921	0.048	0.0772	0.886
RA	9q33.1	116270000	116260000	0.104	0.1245	0.829	0.378	0.399	0.975	0.00528	0.728	0.941
	10q25.2	114080000	114070000	0.041	0.0435	0.828	0.0582	0.129	0.889	0.0259	0.0316	0.947
HT	5q13.2	71873000	71847000	0.1065	0.101	0.754	0.96	0.2	0.956	0.000124	0.0363	0.943
	22q12.3	34870000	34842000	0.091	0.0885	0.741	0.289	0.344	0.977	0.00147	0.986	0.871
T1D	7p12.1	52059000	52056000	0.065	0.085	0.821	0.75	0.647	0.98	0.000152	0.319	0.955

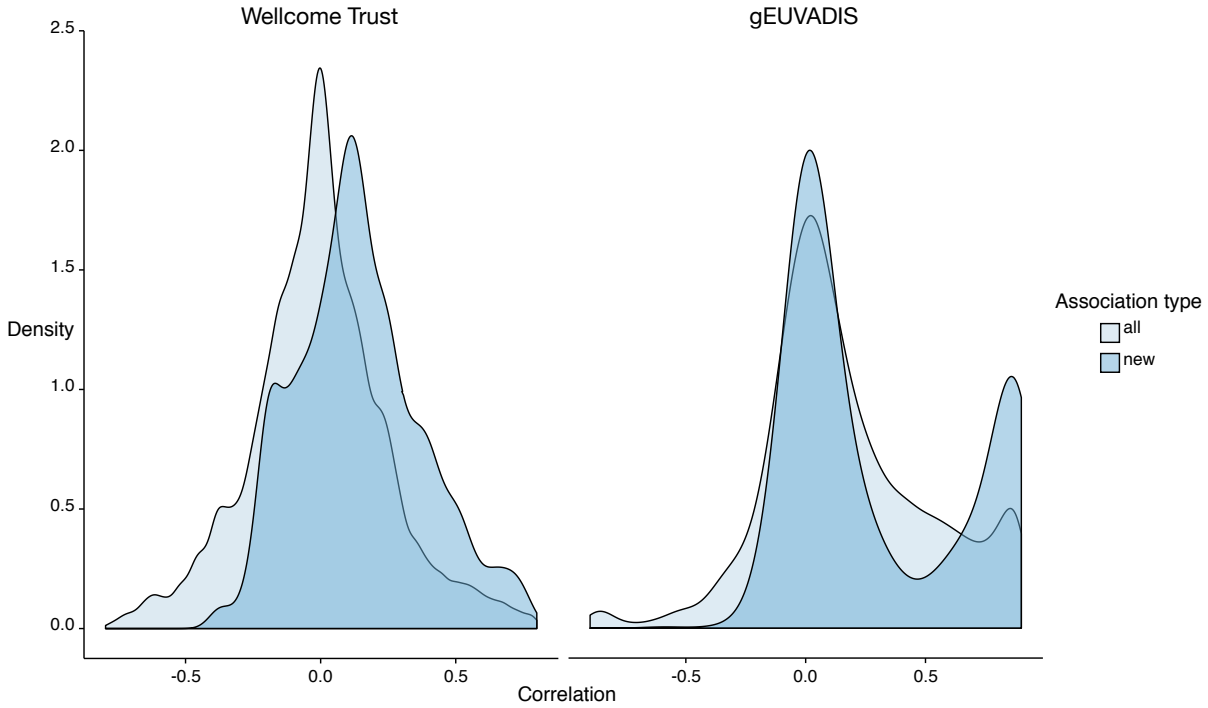

Figure 2.4: (left) Density of the correlation between SNPs in pairs that are genome-wide significant at FWER 5% in the Wellcome Trust dataset, with all pairs of significant SNPs in light blue and just those SNPs discovered using `jester` in dark blue. (right) Density of the correlation between SNPs in the most significant pair for each gene containing an eQTL pair at FDR 5% in the gEUVADIS dataset, with all pairs from genes with significant eQTLs in light blue and just those eQTL's discovered using `jester` in dark blue.

Table 2.8: Joint testing pairs of *cis*-SNPs improves the number of eQTL's detected at FDR 5% by 10.7%. Many of the new genes (row new) discovered using the joint test appear to be linkage masked (row LM), with correlation between the significant SNP pair of above 0.2.

FDR	1%	2%	3%	4%	5%	10%	15%	20%	25%
marg	4528	4936	5240	5458	5641	6432	7072	7600	8123
joint	4884	5376	5726	6015	6248	7206	7947	8625	9277
new	597	698	781	850	908	1113	1263	1431	1565
LM	276	303	334	358	381	473	566	634	701

masking: correlated SNPs with opposite tagging of an un-typed causal, while the remaining two appear to be type 1 linkage masking. Still, we observe increased power due to detection of linkage masking in all aspects of our analysis. The Bulmer effect implies linkage masking should be common; high fitness haplotypes are able to resist selective pressures, and may acquire fitness decreasing mutations without being eliminated from the population. SNPs of this kind are hypothesized as a source of missing heritability by Haig et al. [43] and Lappalainen et al [53] argue that gene expression data support widespread linkage masking due to balancing selection.

We find more evidence for linkage masking in the regulation of gene expression that in the complex disease phenotypes we consider. However, it is not surprising that such effects are more difficult to find in a genotyped cohort. As the effects of such SNPs are already masked, the tagging SNP pair must be in tight LD with the causal SNP pair to prevent severe power loss. On top of this, the tagging SNP pair must be highly correlated to achieve the increased signal necessary to find linkage masked SNP pairs. Hemani et al. [45] make a similar argument, showing that small reductions in LD can result in dramatic under-estimation of epistatic effects. While linkage masked SNPs are difficult to uncover using standard marginal association methods their signal is included in mixed-model SNP heritability estimates. These SNPs are therefore a source of *hidden* heritability: the difference between the phenotypic variance explained by genome-wide significant associations and the phenotypic variance explained by genotyped variation. This is in contrast to the *missing* heritability, the difference between the variance explained by genotyped variation and the total narrow-sense heritability, which is unaffected by linkage masking. Our result narrows the hidden heritability gap by discovering new associations which increase the variance explained due to statistically significant associations.

Our approach is not without drawbacks. When causal variants are sparse we see a reduction in power due to the degree of freedom and multiple test correction which only reaps benefits in the presence of multiple signals. While our results indicate this is a less common situation, there are two loci discovered by a marginal GWAS but not by *Jester* (Table 2.3). Additionally, it can be difficult to distinguish heavy linkage masking from genotyping error. In our analysis of the WT disease associations, we used a window size of 100 SNPs, chosen based on results from Han et al. [44] A logical question is whether or not a larger or smaller window size would lead to more results. We repeated the analysis with a window size of 50 and a correlation cutoff of 0.04, and discovered the same number of loci and slightly more SNPs than presented here (not shown).

The relationship between joint testing and set testing [60, 50] remains an interesting avenue for further investigation. Set and joint testing are similar in that they 1) both improve power to detect associations in the presence of multiple causal variants, 2) both are susceptible to false positives due to genotyping error and population structure and 3) both require an explicit estimation of the multiple testing correction to maintain the desired FWER. We consider one recent set test, FaST-LMM-set, which has shown desirable properties with respect to previous set tests [15]. FaST-LMM-set without a background kernel is equivalent to a likelihood ratio test of the SNPs in the set against a null model, while our test is a

likelihood ratio test of pairs of SNPs. We used FaST-LMM-set to analyze the WTCCC dataset using 100-SNP windows, the same size used in our analysis, but found concerning levels of inflation in the test statistics that we were unable to resolve (Figure 2.4). In addition, set tests currently require expensive permutation tests to control the FWER making genome-wide application computationally intensive [15]. We view approximating the MTC of set testing without resorting to permutations as an interesting open problem, and believe that it may be possible to extend the MVN framework to set tests in the same way we have done for joint tests. We hope to explore this connection thoroughly in future work.

Yang et al. [125] propose a mathematically similar approach to ours, determining joint test statistics from marginal summary test statistics (albeit only at genome-wide significant marginal loci), while using an external reference panel to estimate the pairwise LD. This approach in combination with our multiple hypothesis correction threshold could provide a way to apply our local method without access to genotype data. We caution that our proposed imputation based genotyping error correction method will not be applicable here and thus high LD SNPs should be avoided in such an analysis. Furthermore, the variance of the correlation coefficient estimates can be large even when many hundreds of individuals are available in the reference panel [128] which could lead to false positives. While the reference panel correlations may still be relatively accurate for controls from the same population, case individuals are more likely to harbor many disease-associated mutations and thus will not match the reference panels as well [130, 129]. Even with these caveats, however, the vast gain in power possible with the increased sample size of summary data makes this a tempting proposition, and we have implemented this method in our software package.

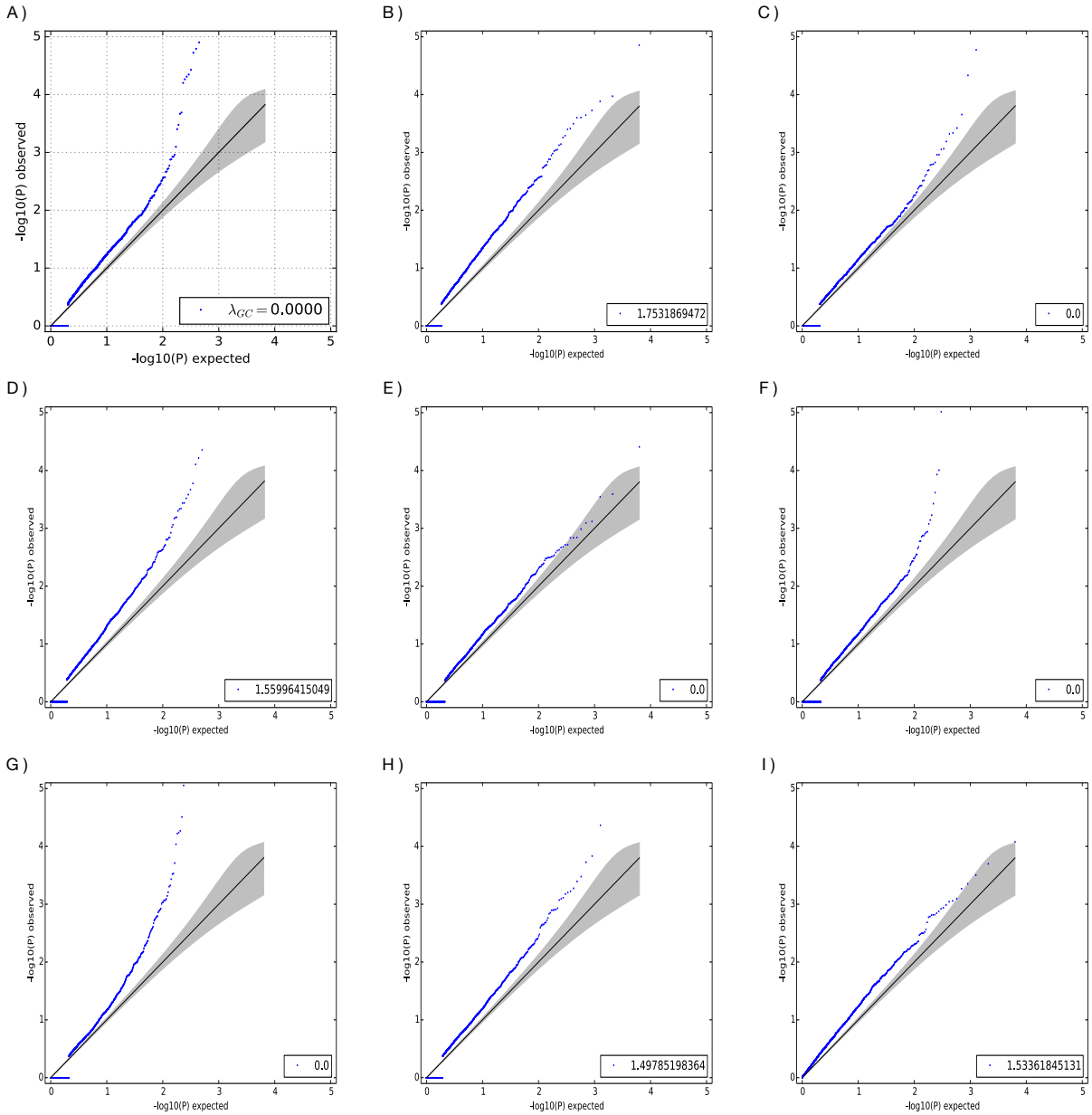

Figure 2.5: Q-Q plots of the results of FaST-LMM on the WTCCC dataset. (A) Control-control analysis, 15.6% of sets have a p-value of less than 0.10. (B) Bipolar disorder, 18.4% of sets have a p-value of less than 0.10. (C) Coronary artery disease, 13.4% of sets have a p-value of less than 0.10. (D) Crohn's disease, 16.5% of sets have a p-value of less than 0.10. (E) Hypertension, 13.4% of sets have a p-value of less than 0.10. (F) Rheumatoid arthritis, 14.2% of sets have a p-value of less than 0.10. (G) Type-1 diabetes, 13.7% of sets have a p-value of less than 0.10. (H) Type-2 diabetes, 14.7% of sets have a p-value of less than 0.10. (I) Bipolar disorder with leave-one-chromosome-out GRM background kernel, 15.4% of sets have a p-value of less than 0.10. In all cases we used 100-SNP sets. In (A) through (H), we used a likelihood ratio test with 10 permutations per set and no background kernel. In (I), we used the `sc_davies` score test to improve speed with the background kernel present. In some cases, λ_{GC} is 0 because permutation tests result in more than half of the sets having a p-value of 1.0.

Chapter 3

Transethnic genetic correlations from summary statistics

3.1 Introduction

Many complex human phenotypes vary dramatically in their distributions between populations due to a combination of genetic and environmental differences. For example, northern Europeans are on average taller than southern Europeans [97] and African Americans have an increased rate of hypertension relative to European Americans [10]. The genetic contribution to population phenotypic differentiation is driven by differences in causal allele frequencies, effect sizes, and genetic architectures. Understanding the root causes of phenotypic differences worldwide has profound implications for biomedical and clinical practice in diverse populations, the transferability of epidemiological results, aiding multi-ethnic disease mapping [23, 74], assessing the contribution of non-additive and rare variant effects, and modeling the genetic architecture of complex traits. In this work we consider a central question in the global study of phenotype: do genetic variants have the same phenotypic effects in different populations?

While the vast majority of GWAS have been conducted in European populations [12], the growing number of non-European and multi-ethnic studies [2, 80, 74] provide an opportunity to study genetic effect distributions across populations. For example, one recent study used mixed-model based methods to show that the genome-wide genetic correlation of schizophrenia between European and African Americans is nonzero [14]. While powerful, computational costs and privacy concerns limit the utility of genotype-based methods. In this work, we make two significant contributions to studies of transethnic genetic correlation. First, we expand the definition of genetic correlation to better account for a transethnic context. Second, we develop an approach to estimating genetic correlation across populations that uses only summary level GWAS data. Similar to other recent summary statistics based methods [83, 57, 125, 86, 47, 46, 52, 8, 7, 34, 84, 123] our approach supplements summary association data with linkage disequilibrium (LD) information from external reference panels,

avoids privacy concerns, and is scalable to hundreds of thousands of individuals and millions of markers. Unlike traditional approaches that focus on the similarity of GWAS results [49, 135, 33, 19, 118] we utilize the entire spectrum of GWAS associations while accounting for LD in order to avoid filtering correlated SNPs.

In a single population, the genetic correlation of two phenotypes is defined as the correlation coefficient of SNP effect sizes [55, 7]. In multiple populations, differences in allele frequency motivate multiple possible definitions of genetic correlation. Here we consider both the correlation of allele effect sizes as well as the correlation of allelic impact, which takes into account the frequency of the variant in the population: a variant may have a much higher effect size but much lower frequency in one population. Therefore, we define the transethnic genetic effect correlation (ρ_{ge}, previously defined by Lee et al [55] and implemented in GCTA) as the correlation coefficient of the per-allele SNP effect sizes, and the transethnic genetic impact correlation (ρ_{gi}) as the correlation coefficient of the population-specific allele variance normalized SNP effect sizes.

Intuitively, the genetic effect correlation measures the extent to which the same variant has the same phenotypic change, while the genetic impact correlation gives more weight to common alleles than rare ones separately in each population. For example, if the effect sizes are the same in each population $\rho_{ge} = 1$ but $\rho_{gi} < 1$ because of allele frequency differences between the populations. In this case $\rho_{gi} < \rho_{ge}$ however the opposite can also be true. If rare alleles have larger effect sizes and there are many alleles common in study one but rare in study two, then ρ_{gi} will be greater than ρ_{ge}. In this case, the differences in effect sizes are mitigated by corresponding differences in allele frequency. While other definitions of the genetic correlation are possible (see discussion), these quantities capture two important questions about the study of disease in multiple populations: to what extent do the same mutations in multiple populations differ in their phenotypic effects and to what extent are these differences mitigated or exacerbated by differences in allele frequency?

To estimate genetic correlation, we take a Bayesian approach wherein we assume genotypes are drawn separately from within each population and effects sizes have a normal prior (the infinitesimal model [31]). While unlikely to represent reality, this model has been used successfully in practice [64, 7, 34, 124, 14]. The infinitesimal assumption yields a multivariate normal distribution on the observed test statistics (Z-scores), which is a function of the heritability and genetic correlation. Rather than pruning SNPs in LD [83, 101, 113], this allows us to explicitly model the resulting inflation of Z-scores. We then maximize an approximate weighted likelihood function to find the heritability and genetic correlation. This method is implemented in a python package called `popcorn`. Though derived for quantitative phenotypes, `popcorn` extends easily to binary phenotypes under the liability threshold model. We show via extensive simulation that `popcorn` produces unbiased estimates of the genetic correlation and the population specific heritabilities, with a standard error that decreases as the number of SNPs and individuals in the studies increases. Furthermore, we show that our approach is robust to violations of the infinitesimal assumption.

We apply `popcorn` to European and Yoruban gene expression data [106] as well as GWAS summary statistics from European and East Asian rheumatoid arthritis and type-two dia-

betes cohorts [75, 21]. Our analysis of gEUVADIS shows that our summary statistic based estimator is concordant with the mixed model based estimator. We find that the mean transethnic genetic correlation across all genes is low (ρ_{ge}= 0.320 (0.009)), but increases substantially when the gene is highly heritable in both populations (ρ_{ge}= 0.772 (0.017)). We find the genetic effect correlation in RA and T2D to be 0.463 (0.058) and 0.621 (0.088), respectively.

Across all phenotypes considered, we overwhelmingly find that the transethnic genetic correlation is significantly less than one. There are many phenomena that may contribute to this, including: untyped and unimputed, possibly rare variants linked to observed SNPs; gene-gene interactions or dominance effects, gene-environment interactions, including epigenetic effects, that are differential between populations; and differences in sub-phenotype composition. Our results therefore show that these phenomena significantly alter the effect sizes of SNPs common to both populations, but cannot differentiate between them on the basis of this analysis. Furthermore, our finding that effects differ between populations indicates that GWAS results may not transfer between populations, necessitating increased study of disease in multiple populations to gain insight into differences in genetic architecture.

3.2 Methods

Our method takes as input summary association statistics from two studies of a phenotype in two different populations, along with two sets of reference genotypes each matching one of the populations in the study. Our method has two steps: first, we estimate the diagonal elements of the LD matrix products $\Sigma_1^2, \Sigma_2^2, \Sigma_1\Sigma_2$, then using these estimates we find the maximum likelihood values and estimate standard errors of the parameters of interest: h_1^2, h_2^2 and ρ_{ge} or ρ_{gi}. The details follow.

Consider two GWAS of a phenotype conducted in different populations populations, A and B. Assume we have N_A individuals genotyped or imputed to M SNPs in study A and N_B individuals genotyped or imputed to M SNPs in study B. Let G_A, G_B and Y_A, Y_B be the matrices of mean-centered genotypes and phenotypes of the individuals in study A and B, respectively, with f_A, f_B the allele frequencies of the M SNPs common to both populations. Assuming Hardy-Weinberg equilibrium, the allele variances are $\sigma_A^2 = 2f_A(1-f_A), \sigma_B^2 = 2f_B(1-f_B)$. Let β_A, β_B be the (unobserved) per-allele effect sizes for each SNP in studies A and B, respectively. Define the *genetic impact correlation* $\rho_{gi} = \mathrm{Cor}(\sqrt{\sigma_A^2}\beta_A, \sqrt{\sigma_B^2}\beta_B)$ and the *genetic effect correlation* $\rho_{ge} = \mathrm{Cor}(\beta_A, \beta_B)$. We present a maximum likelihood framework for estimating the heritability of the phenotype in study A and it's standard error, the heritability of the phenotype in study B and it's standard error, and the genetic effect and impact correlation of the phenotype between the studies and it's standard error given only the summary statistics Z_A, Z_B and reference genotypes R_A, R_B representing the populations in the studies. We assume that genotypes are drawn randomly from populations with expected correlation matrices Σ_A (and similarly for study B), and that every SNP is causal with a normally distributed effects size (though this assumption is not necessary in

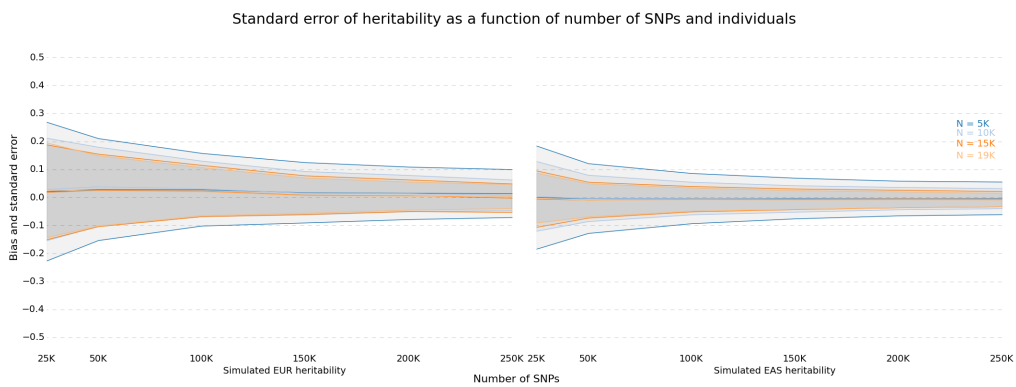

Figure 3.1: Bias and standard error of the heritability estimator in `popcorn` as the number of SNPs M and number of individuals N varies.. All simulations conducted using simulated phenotypes with $h_1^2 = 0.5$, $h_2^2 = 0.5$, $\rho_{gi,e} = 0.5$ and simulated European (EUR) and East Asian (EAS) genotypes generated with HapGen2.

practice, see Figure 3.1).

Genetic impact correlation

Let $X_A = \frac{G_A}{\sqrt{\sigma_A^2}}$ (and similarly for study 2) be normalized genotype matrices. We consider the standard linear model for generation of the phenotypes, where

$$Y_A = X_A\beta_A + \epsilon_A$$
$$Y_B = X_B\beta_B + \epsilon_B$$

For convenience of notation let $h_{ix}^2 = \rho_{gi}\sqrt{h_A^2 h_B^2}$. We assume the SNP effects follow the infinitesimal model, where every SNP has an effect size drawn from the normal distribution, and that the residuals are independent for each individual and normally distributed:

$$\begin{pmatrix} \beta_A \\ \beta_B \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \frac{1}{M}\begin{bmatrix} h_A^2\mathbb{I}_M & h_{ix}^2\mathbb{I}_M \\ h_{ix}^2\mathbb{I}_M & h_B^2\mathbb{I}_M \end{bmatrix}\right) \tag{3.1}$$

$$\begin{pmatrix} \epsilon_A \\ \epsilon_B \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} (1-h_A^2)\mathbb{I}_M & 0 \\ 0 & (1-h_B^2)\mathbb{I}_M \end{bmatrix}\right) \tag{3.2}$$

where h_A^2, h_B^2 are the heritability of the disease in study one and two, respectively, and ρ_{gi} is the genetic impact correlation.

Using the above model, we compute the distribution of the observed Z scores as a function of the reference panel correlations and the model parameters $(h_A^2, h_B^2, \rho_{gi})$. Given a distribution for Z and an observation of Z we can then choose parameters which give the highest probability of observing Z. First, we compute the distribution of Z. It is well known

that the Z-scores of a linear regression are normally distributed given β when the sample size is large enough. Since $\mathbb{P}(Z) \propto \mathbb{P}(Z|\beta)\mathbb{P}(\beta)$ and the product of normal distributions is normal, we only need to compute the unconditional mean and variance of Z to know its distribution. Specifically, let $Z = [Z_A^\top, Z_B^\top]^\top$. Let $Z \sim \mathcal{N}\left(\begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix}, \begin{bmatrix} C_{11} & C_{12} \\ C_{12}^\top & C_{22} \end{bmatrix}\right)$ From Chapter 1, we know that the mean and within-population variance are

$$\begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix} = \mathbb{E}\begin{bmatrix} Z_A \\ Z_B \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$
$$C_{11} = \mathbb{E}[Z_A Z_A^\top] = \Sigma_A + h_A^2 \frac{N_A+1}{M}\Sigma_A^2$$
$$C_{22} = \mathbb{E}[Z_B Z_B^\top] = \Sigma_B + h_B^2 \frac{N_B+1}{M}\Sigma_A^2$$

The between-population variance C_{12} is

$$C_{12} = \mathbb{E}[Z_A Z_B^\top] = \frac{1}{\sqrt{N_A N_B}}\mathcal{E}\left[X_A^\top Y_A Y_B^\top X_B\right] \tag{3.3}$$
$$= \frac{1}{\sqrt{N_A N_B}}\mathbb{E}_{X_A,X_B}\left[X_A^\top \mathbb{E}\left[Y_A Y_B^\top | X_A, X_B\right] X_B\right] \tag{3.4}$$
$$= \frac{h_{ix}}{M\sqrt{N_A N_B}}\mathbb{E}\left[X_A^\top X_A X_B^\top X_B\right] \tag{3.5}$$
$$= h_{ix}\frac{\sqrt{N_A N_B}}{M}\Sigma_A \Sigma_B \tag{3.6}$$

Genetic effect correlation

Let $h_{ex} = \rho_{ge}\sqrt{h_A^2 h_B^2}$. We modify the procedure above to use mean-centered instead of normalized genotype matrices and model the distribution of the effect sizes as

$$\begin{pmatrix} \beta_A \\ \beta_B \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{h_A^2}{\|\sigma_A^2\|_1}\mathbb{I}_M & \frac{h_{ex}}{\sqrt{\|\sigma_A^2\|_1\|\sigma_B^2\|_1}}\mathbb{I}_M \\ \frac{h_{ex}}{\sqrt{\|\sigma_A^2\|_1\|\sigma_B^2\|_1}}\mathbb{I}_M & \frac{h_B^2}{\|\sigma_B^2\|_1}\mathbb{I}_M \end{bmatrix}\right) \tag{3.7}$$

Notice that a linear model with effects sizes acting on un-normalized genotypes is the same as a linear model with effect sizes acting on normalized genotypes under the substitution $\beta_{A,B} \to \sqrt{\sigma_{A,B}^2}\beta_{A,B}$. Therefore the covariance of Z-scores on the per allele scale can be immediately inferred from the prior derivation

$$C = \mathrm{Var}(Z) = \begin{bmatrix} \Sigma_A + \frac{N_A+1}{\|\sigma_A^2\|_1}h_A^2\Sigma_A\sigma_A^2\Sigma_A & h_{gx}^2\frac{\sqrt{N_A N_B}}{\sqrt{\|\sigma_A^2\|_1\|\sigma_B^2\|_1}}\Sigma_A\sqrt{\sigma_A^2\sigma_B^2}\Sigma_B \\ h_{gx}^2\frac{\sqrt{N_A N_B}}{\sqrt{\|\sigma_A^2\|_1\|\sigma_B^2\|_1}}\Sigma_B\sqrt{\sigma_B^2\sigma_A^2}\Sigma_A & \Sigma_B + \frac{N_B+1}{\|\sigma_B^2\|_1}h_B^2\Sigma_B\sigma_B^2\Sigma_B \end{bmatrix}$$

Approximate maximum likelihood estimation

We approximately optimize the above likelihood as follows: first we find h_A^2 and h_B^2 by maximizing the likelihood corresponding to C_{11} and C_{22}, then we find ρ_{gi} or ρ_{ge} by maximizing the likelihood corresponding to C_{12}:

$$l(h_A^2|Z_A,\Sigma,\sigma) \approx -\sum_{i=1}^{M} w_{11i}\left(\ln(C_{11ii}) + \frac{Z_{1i}^2}{C_{11ii}}\right)$$

$$l(h_B^2|Z_B,\Sigma,\sigma) \approx -\sum_{i=1}^{M} w_{22i}\left(\ln(C_{22ii}) + \frac{Z_{2i}^2}{C_{22ii}}\right)$$

$$l(\rho_{g\{i,e\}}|Z,\hat{h_A^2},\hat{h_B^2},\Sigma,\sigma) \approx -\sum_{i=1}^{M} w_{12i}\left(\ln(C_{12ii}) + \frac{Z_{1i}Z_{2i}}{C_{12ii}}\right)$$

Because we are discarding between-SNP covariance information ($\mathrm{Cov}(Z_{Ai}, Z_{Aj})$), highly correlated SNPs will be overcounted in our approximate likelihood. As a simple example, notice that two SNPs in perfect LD will each contribute identical terms to the approximate likelihood, and therefore should be downweighted by a factor of 1/2. The extent to which SNP i is over-counted is exactly the i'th entry in it's corresponding LD-matrix product. Therefore we let $w_{jki}^{gi} = 1/\left(\Sigma_j\Sigma_k\right)_{ii}$ and $w_{jki}^{ge} = 1/\left(\Sigma_j\sqrt{\sigma_j^2\sigma_k^2}\Sigma_k\right)_{ii}$ to reduce the variance in our estimates of the parameters h_A^2, h_B^2, ρ_{gi} and ρ_{ge}.

Furthermore, rather than compute the full products Σ_1^2, Σ_2^2 and $\Sigma_1\Sigma_2$ over all M SNPs in the genome, we choose a window size W and approximate the product by $(\Sigma_a\Sigma_b)_{ii} = \sum_{w=i-W}^{w=i+W} r_{aiw}r_{biw}$. Though maximum likelihood estimation admits a straightforward estimate of the standard error via the fisher information, we found these estimates to be inaccurate in practice. Instead, we use block jackknife with block size equal to $\min(100, \frac{M}{200})$ SNPs to ensure that blocks are large enough to remove residual correlations. These optimizations are similar to those employed by LD score regression [7].

Out of population prediction of phenotypic values

Consider using the results of a GWAS with perfect power in population 2 to predict the phenotypic values of a set of individuals from population 1. This defines the upper limit of the correlation of true and predicted phenotypic values. Let the true values of the effects sizes in population 2 be β_B. Let the true phenotypes in population 1 be $Y = X_A\beta_A + \epsilon_A$ while the predicted phenotypes are $P = X_A\beta_B$. We are interested in the correlaiton of the predicted and true phenotypes $\rho_{YP}^{MAX} = \mathrm{Cor}(Y, P)$. Notice that, given X, the true and predicted phenotype of each individual is an affine transformation of a multivariate normal random variable (β)

$$\begin{bmatrix} Y_i \\ P_i \end{bmatrix} = \begin{bmatrix} X_{(i)} & 0_M \\ 0_M & X_{(i)} \end{bmatrix} \begin{bmatrix} \beta_A \\ \beta_B \end{bmatrix} + \begin{bmatrix} \epsilon_i \\ 0 \end{bmatrix}$$

and therefore (Y_i, P_i) for individual i is multivariate normal with expected covariance matrix

$$\begin{aligned}
\mathbb{E}_X\left[\mathrm{Cov}(Y_i, P_i)\right] &= \mathbb{E}_X \begin{bmatrix} X_{(i)} & 0_M \\ 0_M & X_{(i)} \end{bmatrix} \begin{bmatrix} \frac{1}{\|\sigma_A^2\|_1}\mathbb{I}_M & \frac{h_{ex}}{\sqrt{\|\sigma_A^2\|_1\|\sigma_B^2\|_1}}\mathbb{I}_M \\ \frac{h_{ex}}{\sqrt{\|\sigma_A^2\|_1\|\sigma_B^2\|_1}}\mathbb{I}_M & \frac{h_2^2}{\|\sigma_B^2\|_1}\mathbb{I}_M \end{bmatrix} \begin{bmatrix} X_{(i)} & 0_M \\ 0_M & X_{(i)} \end{bmatrix}^\top \\
&= \mathbb{E}_X \begin{bmatrix} \frac{\sum_m X_{im}^2}{\|\sigma_A^2\|_1} & \frac{h_{ex}\sum_m X_{im}^2}{\sqrt{\|\sigma_A^2\|_1\|\sigma_B^2\|_1}} \\ \frac{h_{ex}\sum_m X_{im}^2}{\sqrt{\|\sigma_1^2\|_1\|\sigma_B^2\|_1}} & \frac{h_2^2\sum_m X_{im}^2}{\|\sigma_B^2\|_1} \end{bmatrix} \\
&= \begin{bmatrix} 1 & h_{ex}\sqrt{\frac{\|\sigma_A^2\|_1}{\|\sigma_B^2\|_1}} \\ h_{ex}\sqrt{\frac{\|\sigma_A^2\|_1}{\|\sigma_B^2\|_1}} & h_B^2\frac{\|\sigma_A^2\|_1}{\|\sigma_B^2\|_1} \end{bmatrix}
\end{aligned}$$

and therefore the expected correlation $\mathbb{E}[\mathrm{Cor}(Y_i, P_i)]$ is $\frac{h_{ex}}{\sqrt{h_B^2}}\sqrt{\frac{\|\sigma_A^2\|_1}{\|\sigma_B^2\|_1}\frac{\|\sigma_B^2\|_1}{\|\sigma_A^2\|_1}} = \rho_{ge}\sqrt{h_A^2}$. The expected population correlation tends to the sample correlation as the number of samples increases, therefore

$$\rho_{YP}^{MAX} = \mathrm{Cor}(Y, P) \to \rho_{ge}\sqrt{h_A^2} \tag{3.8}$$

as $N \to \infty$

3.3 Results

Simulated genotypes and simulated phenotypes

In order to verify that `popcorn` yields an unbiased estimate of the heritabilities (h_A^2, h_B^2) and genetic correlations (ρ_{ge}, ρ_{gi}), we applied `popcorn` to summary statistics from simulated GWAS. We simulated 50,000 European-like (EUR) and 50,000 East Asian-like (EAS) individuals at 248,953 SNPs from chromosomes 1-3 with allele frequency above 1% in both European and East Asian HapMap3 populations with HapGen2 [105]. HapGen2 implements a haplotype recombination with mutation model that results in excess local relatedness among the simulated individuals. To account for this local structure, we used Plink2 [18] to filter individuals with genetic relatedness above 0.05, resulting in 4499 EUR-like individuals and 4837 EAS-like individuals. From these simulated individuals, 500 per population were chosen uniformly at random to serve as an external reference panel for estimating Σ_A and Σ_B.

In each simulation effect sizes were drawn from a ?spike and slab? model, where $\beta_{1i}, \beta_{2i} \sim \mathcal{N}\left(0, \begin{bmatrix} h_A^2 & \rho_{ge}\sqrt{h_A^2 h_B^2} \\ \rho_{ge}\sqrt{h_A^2 h_B^2} & h_2^2 \end{bmatrix}\right)$ with probability p and $\beta_{Ai}, \beta_{Bi} = (0, 0)$ with probability $1 - p$. ρ_{gi} was analytically computed from the simulated effect sizes and allele frequencies in the simulated reference genotypes. Quantitative phenotypes were generated under a linear model with i.i.d. noise and normalized to have mean 0 and variance 1, while binary phenotypes were generated under a liability threshold model.

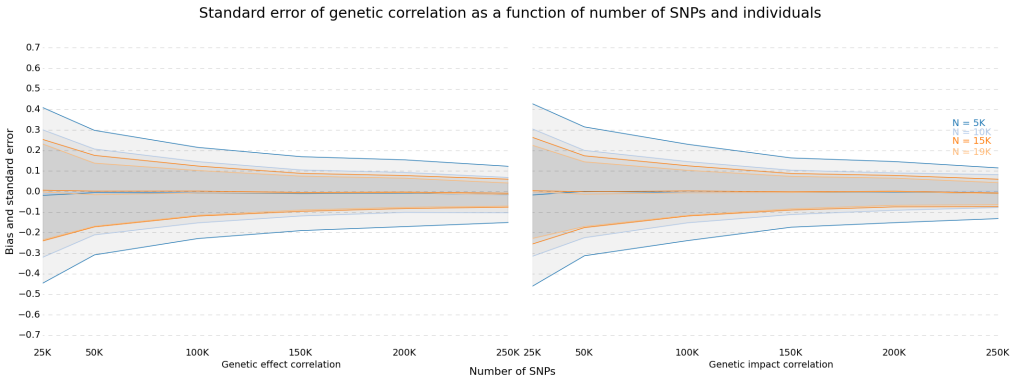

Figure 3.2: True and estimated genetic impact and effect correlation. All simulations conducted with simulated EUR and EAS heritability of 0.5 using 4499 simulated EUR and 4836 simulated EAS individuals at 248,953 SNPs.

Table 3.1: Average heritability and genetic correlation over 1000 simulations with varying levels of ascertainment. All simulations contained exactly N cases and controls for a study prevalence of 0.5. Phenotypes were simulated with liability scale heritability of 0.3 for both phenotypes and genetic correlation of 0.3.

Prevalence	N	$\hat{h}_1$	$\hat{h}_2$	$\hat{\rho}_g$
0.03	1000	0.31	0.31	0.27
0.05	1700	0.30	0.31	0.30
0.1	3400	0.31	0.30	0.29
0.25	5000	0.31	0.30	0.30

We varied h_A^2, h_B^2, ρ_{ge}, and ρ_{gi}, as well as the number of individuals in each study (N_A, N_B), the number of SNPs (M), the population prevalence K, and proportion of causal variants (p) in the simulated GWAS and generated summary statistics for each study. The results shown in Figure 3.2 and Figure 3.1 demonstrate that the estimators are nearly unbiased as the genetic correlation and heritabilities vary. Furthermore, by varying the proportion of causal variants p we show that our estimator is robust to violations of the infinitesimal assumption (Figure 3.3). In figure 3.4, we show that the standard error of the estimator decreases as the number of SNPs and individuals in the study increases. Finally, we show in Table 3.1 that our estimates of the heritability of liability in case control studies are nearly unbiased.

Simulations with nonstandard disease models

Our approach, as well as genotype-based methods such as GCTA, makes assumptions about the genetic architecture of complex traits [102]. Previous work has shown that violations of

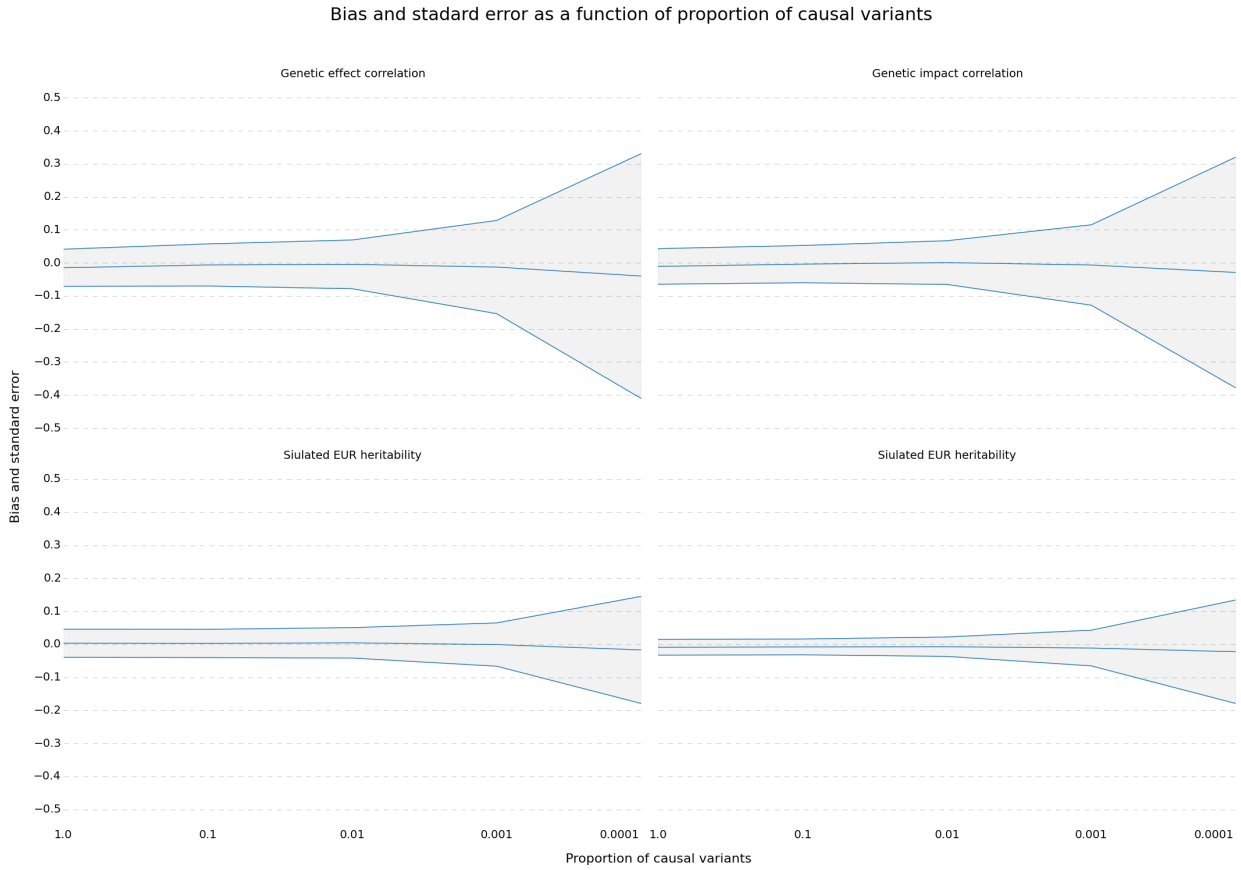

Figure 3.3: Bias and standard error as the proportion of causal variants is decreased from 1.0 (all variants causal, the infinitesimal model) to 0.0001 (one in ten thousand variants causal, or approximately 25 total causals). All simulations conducted using simulated phenotypes with $h_1^2 = 0.5$, $h_2^2 = 0.5$, $\rho_{gi,e} = 0.5$ and simulated European (EUR) and East Asian (EAS) genotypes generated with HapGen2.

these assumptions can lead to bias in heritability estimation, therefore we sought to quantify the extent that this bias may effect our estimates. We simulated phenotypes under six different disease models. Independent: effect size independent of allele frequency. Inverse: effect size inversely proportional to allele frequency. Rare: only SNPs with allele frequency under 10% affect the trait. Common: only SNPs with allele frequency between 40% and 50% affect the trait. Difference: effect size proportional to difference in allele frequency. Adversarial: difference model with sign of beta set to increase the phenotype in the population where the allele is most common. Additional genetic architectures are possible, including ones where effect sizes are not a direct function of MAF [127].

We simulated phenotypes using genotypes with allele frequency above 1% or 5% and compared the true and estimated genetic impact and effect correlation among all models

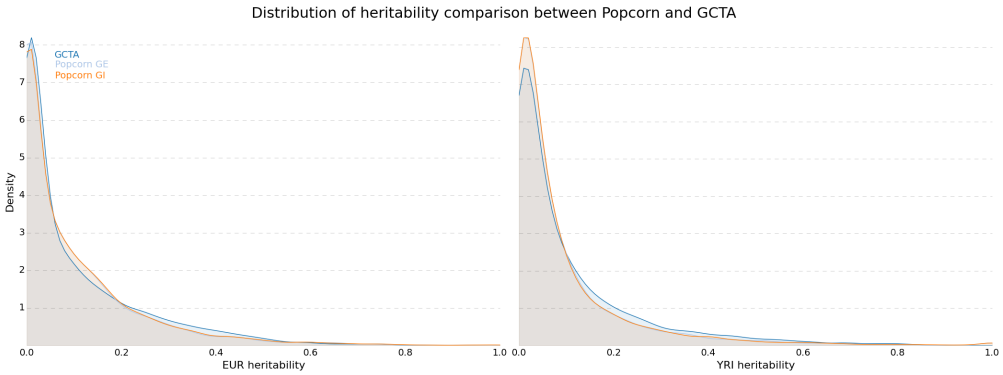

Figure 3.4: Kernel density estimate (KDE) comparison between `popcorn` and GCTA as heritability estimators. Density was computed using the scipy statistics package `gaussian_kde` function on the set of heritability estimates.

Table 3.2: True and estimated values of the genetic impact and effect correlation in simulated EUR-like and EAS-like genotypes. Results are the average of 100 simulations with phenotype heritability of 0.5 in each population.

	MAF >0.01				MAF >0.05			
Model	ρ_{ge}	ρ_{gi}	$\hat{\rho}_{ge}$	$\hat{\rho}_{gi}$	ρ_{ge}	ρ_{gi}	$\hat{\rho}_{ge}$	$\hat{\rho}_{gi}$
Independent	0.500	0.478	0.500	0.460	0.500	0.488	0.509	0.469
Inverse	0.431	0.500	0.567	0.496	0.479	0.500	0.555	0.482
Rare	0.500	0.467	0.382	0.863	0.500	0.496	0.998	0.756
Common	0.500	0.500	0.522	0.493	0.500	0.500	0.502	0.496
Difference	0.500	0.416	0.354	0.435	0.500	0.461	0.410	0.412
Adversarial	0.710	0.604	0.525	0.651	0.714	0.667	0.601	0.675

(Table 3.2). We find that when only SNPs with frequency above 5% in both populations are used, the difference in ρ_{ge} and ρ_{gi} is minimal except in the most adversarial cases. Even in the adversarial model, the true difference is only 7%. Though unlikely to represent reality, the four nonstandard disease models result in substantial bias in our estimators. When SNPs with allele frequency above 1% in both populations are included, the differences are more pronounced. This is because the normalizing constant $1/\sigma$ rapidly increases as the SNP becomes more rare. Indeed, as SNPs become more rare having an accurate disease model becomes increasingly important and therefore we proceed with a 5% MAF cutoff in our analysis of real data and use the notation h_c^2 to refer to the heritability of SNPs with allele frequency above 5% in both populations (the common-SNP heritability). Note, however, that one of the advantages of maximum likelihood estimation in general is that the likelihood can be reformulated to mimic the disease model of interest.

Validation of popcorn using gene expression in gEUVADIS

To further validate our approach, we compared the common-SNP heritability (h_c^2) and genetic correlation estimates of `popcorn` to GCTA in the gEUVADIS dataset for which raw genotypes are publicly available. gEUVADIS consists of RNA-seq data for 464 lymphoblastoid cell line (LCL) samples from five populations in the 1000 genomes project. Of these, 375 are of European ancestry (CEU, FIN, GBR, TSI) and 89 are of African ancestry (YRI). Raw RNA-sequencing reads obtained from the European Nucleotide Archive were aligned to the transcriptome using UCSC annotations matching hg19 coordinates. RSEM was used to estimate the abundances of each annotated isoform and total gene abundance is calculated as the sum of all isoform abundances normalized to one million total counts or transcripts per million (TPM). For eQTL mapping, European and Yoruban samples were analyzed separately. For each population, TPMs were median-normalized to account for differences in sequencing depth in each sample and standardized to mean 0 and variance 1. Of the 29763 total genes, 9350 with $TPM > 2$ in both populations were chosen for this analysis.

For each gene we conducted a cis-eQTL association study at all SNPs within 1 megabase of the gene body with allele frequency above 5% in both populations using 30 principal components as covariates. We found that GCTA and `popcorn` agree on the global distribution of heritability (Figure 3.4) and that GCTA's estimates of genetic correlation have a similar distribution to `popcorn`'s genetic effect (GE) and genetic impact (GI) correlation estimates (Figure 3.5). While the number of SNPs and individuals included in each gene analysis are too small to obtain accurate point estimates of the genetic correlation on a per-gene basis ($N = 464, M = 4279.5$), the large number of genes enables accurate estimation of the global mean heritability and genetic correlation.

Common-SNP heritability and genetic correlation of gene expression in gEUVADIS

We find that the average cis-h_c^2 of the expression of the genes we analyzed was 0.093 (0.002) in EUR and 0.088 (0.002) in YRI. Our estimates are higher than previously reported average cis-heritability estimates of 0.055 in whole blood and 0.057 in adipose [93], which could arise for several reasons. First, we remove 68% of the transcripts that are lowly expressed in either the YRI or EUR data. Second, estimates from RNA-seq analysis of cell lines might not be directly comparable to microarray data from tissue.

The average genetic effect correlation was 0.320 (0.010) while the average genetic impact correlation was 0.313 (0.010). Notably, the genetic correlation increases as the cis-h_c^2 of expression in both populations increases (Figure 3.6). In particular, when the cis-h_c^2 of the gene is at least 0.2 in both populations the genetic effect correlation was 0.772 (0.017) while the genetic impact correlation was 0.753 (0.018).

In order to verify that there were no small-sample size or conditioning biases in our analysis, we analyzed the genetic correlation of simulated phenotypes over the gEUVADIS genotypes. We sampled pairs of heritabilities from the estimated expression heritability dis-

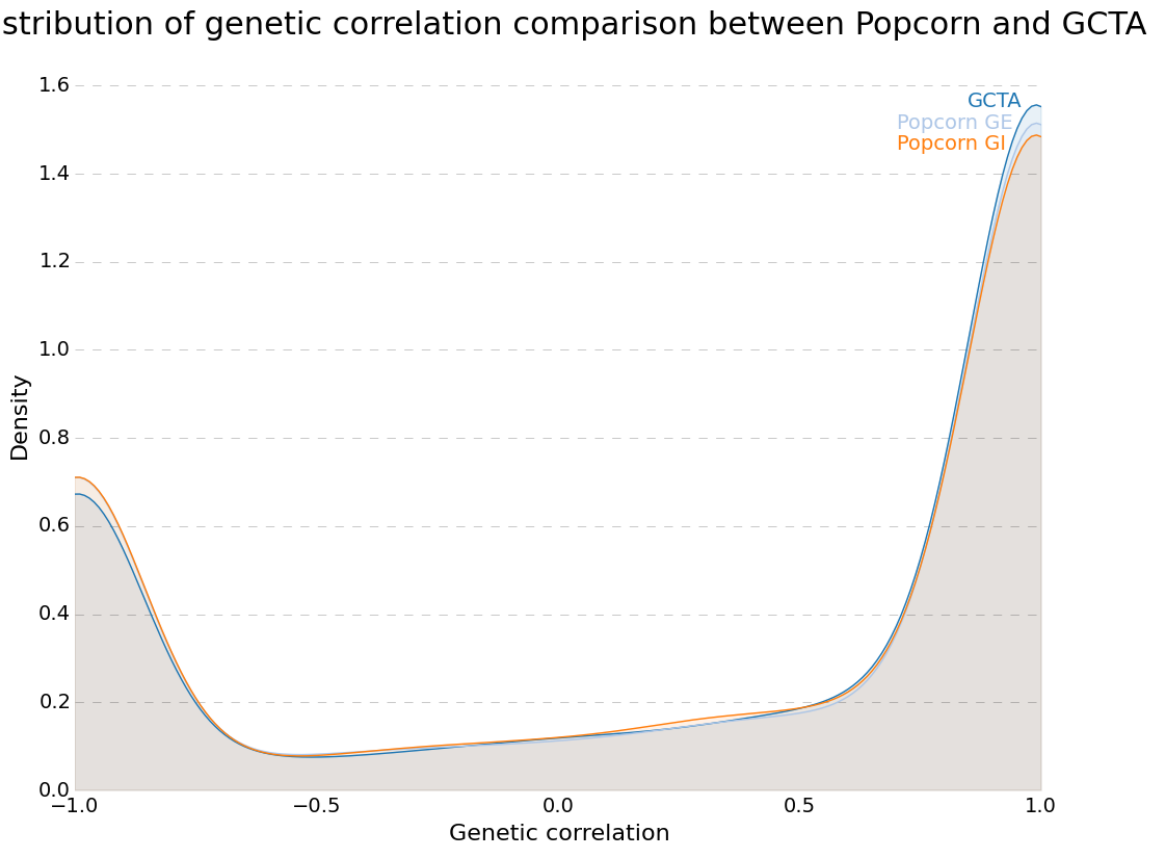

Figure 3.5: Kernel density estimate of genetic correlation comparison between `popcorn` and GCTA. Distribution was computed using a gaussian kde on the set of genetic correlation estimates..

tribution and simulated pairs of phenotypes to have the given heritability and a genetic effect correlation of 0.0 over randomly chosen 4000 base regions from chromosome 1 of the gEUVADIS genotypes. Without conditioning, the average estimated genetic effect correlation was -0.002 (0.003), indicating that the estimator remained unbiased. Furthermore, the average estimated genetic effect correlation was not significantly different from 0.0 conditional on the estimates of heritability being above a certain threshold (Figure 3.7).

We find that while the average genetic correlation is low, the genetic correlation increases with the cis-h_c^2 of the gene, indicating that as cis-genetic regulation of gene expression increases it does so similarly in both YRI and EUR populations. This may help interpret the recent observation that while the global genetic correlation of gene expression across tissues is low [93], cis-eQTL's tend to replicate across tissues [36]. As the presence of a cis-eQTL indicates substantial cis-genetic regulation, an analysis of eQTL replication across tissues is

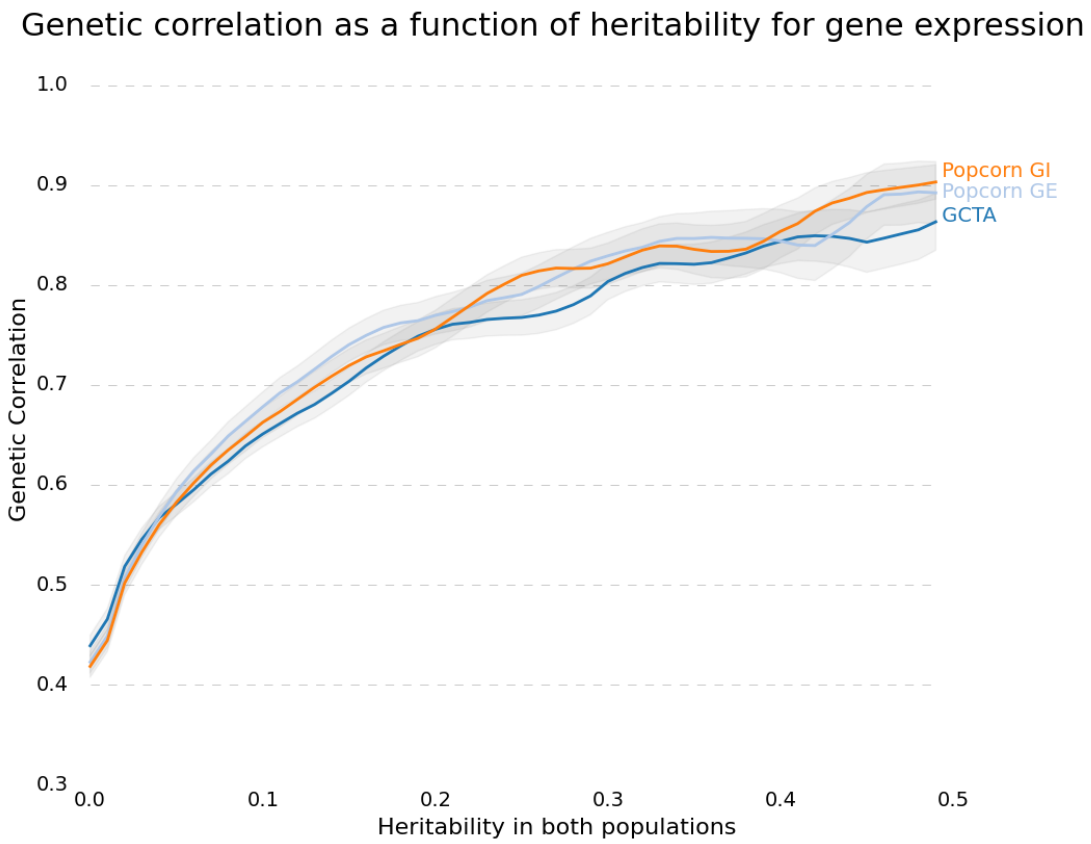

Figure 3.6: Genetic correlation as a function of heritability for gene expression. The mean and standard error of the genetic correlation of the set of genes with h12 and h22 exceeding threshold h in each analysis (y-axis) is plotted against h (x-axis).

implicitly conditioning on the heritability of gene expression being high and therefore may indicate much higher genetic correlation than average.

Summary statistics of RA and T2D

Finally, we sought to examine the transethnic ρ_{gi} and ρ_{ge} in RA and T2D cohorts for which raw genotypes are not available. We obtained summary statistics of GWAS for rheumatoid arthritis and type-2 diabetes conducted in European and East Asian populations. We used genotypes from 504 East Asian and 503 European individuals sequenced as part of the 1000 genomes project as population-specific external reference panels for our EAS and EUR summary statistics, respectively. We removed the MHC region (chromosome 6, 25?35 Mb) from the RA summary statistics. We estimated the common-SNP heritability and genetic

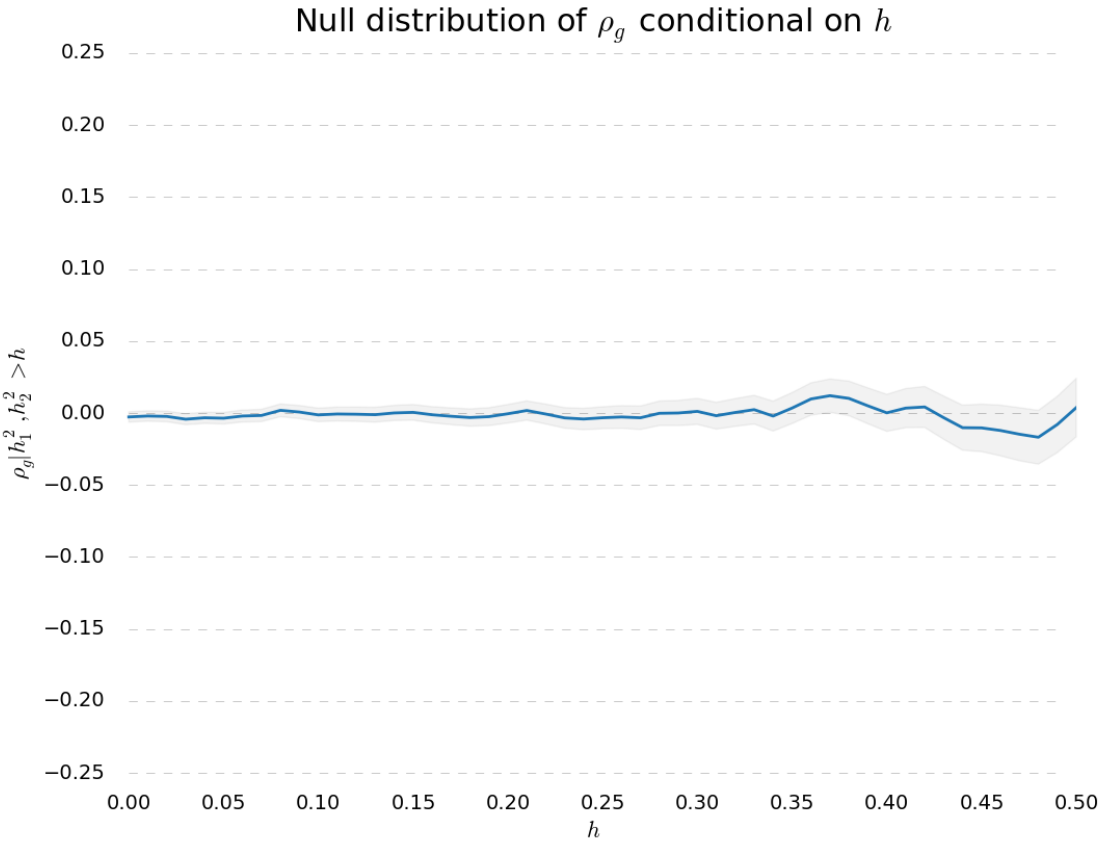

Figure 3.7: Null distribution of the conditional genetic correlation. Phenotypes were generated with heritability randomly sampled from the joint distribution of the gEUVADIS heritability estimates over randomly selected 4000 SNP regions from chromosome 1 of the true EUR and YRI genotypes and genetic correlation of 0. The mean and standard error of the genetic correlation of the set of genes with $\hat{h}_1^2$ and $\hat{h}_2^2$ exceeding threshold h in each analysis (y-axis) is plotted against h (x-axis)

Table 3.3: Heritability and genetic correlation of RA and T2D between EUR and EAS populations. EUR RA data contained 8,875 cases and 29,367 controls for a study prevalence of 0.23. EAS RA data contained 4,873 cases and 17,642 controls for a study prevalence of 0.22. RA disease prevalence was assumed to be 0.5% in both populations. T2D EUR data contained 12171 cases and 56862 controls for a study prevalence of 0.18. T2D EAS data contained 6952 cases and 11865 controls for a study prevalence of 0.37. T2D EUR prevalence was assumed to be 8% while T2D EAS prevalence was assumed to be 9%

		h^2_{EUR} lia	h^2_{EAS} lia	ρ_{ge}	ρ_{gi}
RA	Est. (SE)	0.18 (0.02)	0.22 (0.03)	0.46 (0.06)	0.46 (0.06)
	95% CI	[0.15, 0.21]	[0.16, 0.28]	[0.34, 0.58]	[0.34, 0.58]
	p>0	3.90e-32	1.89e-17	1.37e-15	8.16e-16
	p<1	0.0	3.1e-197	2.53e-20	4.87e-22
T2D	Est. (SE)	0.24 (0.01)	0.11 (0.02)	0.62 (0.09)	0.61 (0.08)
	95% CI	[0.22, 0.26]	[0.07, 0.15]	[0.44, 0.80]	[0.45,0.77]
	p>0	2.41e-77	5.73e-7	1.70e-12	2.85e-13
	p<1	0.0	0.0	1.066e-05	2.06e-06

correlation using 2,539,629 SNPs genotyped or imputed in both RA studies and 1,054,079 SNPs genotyped or imputed in both T2D studies with allele frequency above 5% in 1000 genomes EUR and EAS populations. The h^2_c and genetic correlation estimates are presented in Table 3.3. Our RA h^2_c estimates of 0.177 (0.015) and 0.221 (0.026) for EUR and EAS, respectively, are lower than a previously reported mixed-model based heritability estimates of 0.32 (0.037) in Europeans [103]. Similarly, our T2D h^2_c estimates of 0.242 (0.013) and 0.105 (0.021) for EUR and EAS, respectively, are lower than a previously reported mixed-model based estimate of 0.51 (0.065) in Europeans [103]. We stress that this discrepancy is likely due to the difference between common-SNP heritability h^2_c and total narrow-sense heritability h^2_g. Furthermore, estimates of the heritability of T2D from family studies can vary significantly [70, 79].

We find the genetic effect correlation in RA and T2D to be 0.463 (0.058) and 0.621 (0.088), respectively, while the genetic impact correlation is not significantly different at 0.455 (0.056) and 0.606 (0.083). The transethnic genetic impact and effect correlation for both phenotypes are significantly different from both 1 and 0 (Table 3.3), showing that while there is clear genetic overlap between the phenotypes, the per allele effects sizes differ significantly between the two populations.

3.4 Discussion

We have developed the transethnic genetic effect and genetic impact correlation and provided an estimator for these quantities based only on summary-level GWAS information and

suitable reference panels. We have applied our estimator to several phenotypes: rheumatoid arthritis, type-2 diabetes and gene expression. While the gEUVADIS dataset lacks the power required to make inferences about the genetic correlation of single or small subsets of genes, we can make inferences about the global structure of genetic correlation of gene expression. We find that the global mean genetic correlation is low, but that it increases substantially when the heritability is high in both populations. In all phenotypes analyzed, the genetic correlation is significantly different from both 0 and 1. Our results show that global differences in SNP effect size of complex traits can be large. In contrast, effects sizes of gene expression appear to be more conserved where there is strong genetic regulation.

It is not possible to draw conclusions about polygenic selection from estimates of transethnic genetic correlation. The effects sizes may be identical ($\rho_{ge} = 1$) while polygenic selection acts to change only the allele frequencies. Similarly, the effects sizes may be different ($\rho_{ge} < 1$) without selection. Differences in effects sizes at common SNPs can result from many phenomena. If a gene-gene or gene-environment interaction exists, but only marginal effects are tested, the observed marginal effects will be altered by changes in allele frequency even if the interaction effect is the same in both populations, resulting in decreased genetic correlation. Un-typed and un-imputed variants differentially linked to observed SNPs, as well as differentially tagged rare variants will also contribute, though we expect the latter effect to be small. Another contribution likely comes from variants that are rare in population one but common in population two (and vice versa), which will be filtered in our analysis. While within-locus (dominance) interactions may also play a role [20], the magnitude of this effect has been debated [132]. We emphasize that we cannot differentiate between these effects on the basis of this analysis alone, and further research is required to establish the magnitude of the contribution of each of these effects to inter-population effect size differences.

Estimates of the transethnic genetic correlation are important for several reasons. They may help inform best practices for transethnic meta-analysis, potentially offering improvements over current methods that use F_{st} to cluster populations for analysis [74]. Further, the transethnic genetic correlation constrains the limit of out of sample phenotype predictive power. If the maximum within population correlation of predicted phenotype P to true phenotype Y is $\rho_{YP}^{max} = \sqrt{h_1^2}$, then the maximum out of population correlation is $\rho_{YP}^{max} = \rho_{ge}\sqrt{h_1^2}$ (Methods). Our observation that for RA, T2D, and gene expression the genetic correlation is low shows that out of population phenotypic predictive power is quite low. Similarly, it implies that disease risk assessment in non-Europeans based on current GWAS results may be problematic, necessitating increased study of disease in many populations to gain insight into differences in genetic architecture and improve risk assessment.

While the genetic correlation of multiple phenotypes in one population has a relatively straightforward definition, extending this to multiple populations motivates multiple possible extensions. In this work we have provided estimators for the correlation of genetic effect and genetic impact but other quantities related to the shared genetics of complex traits between populations include the correlation of variance explained $\rho_{ge} = \mathrm{Cor}(\sigma_1^2\beta_1^2, \sigma_2^2\beta_2^2)$ and proportion of shared causal variants between the two populations. Interestingly, while our goal was to construct an estimator that determined the extent of genetic sharing independent

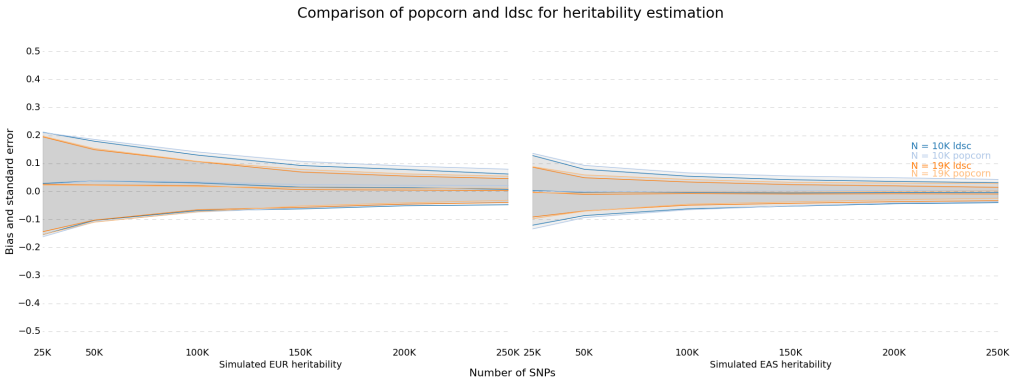

Figure 3.8: Comparison of `popcorn` and ldsc as heritability estimators as the number of SNPs and individuals in each study varies. All simulations conducted using simulated phenotypes with $h_1^2 = 0.5$, $h_2^2 = 0.5$, $\rho_{gi,e} = 0.5$ and simulated European (EUR) and East Asian (EAS) genotypes generated with HapGen2.

of allele frequency, we observe that the correlation of genetic effect and genetic impact are similar. Furthermore, our simulations show that under a random effects model utilizing only SNPs with allele frequency above 5% in both populations the true genetic effect and genetic impact correlation are similar. We conclude that at variants common in both populations, differences in effect size and not allele frequency are driving the transethnic phenotypic differences in these traits.

Our approach to estimating genetic correlation has two major advantages over mixed-model based approaches. First, utilizing summary statistics allows application of the method without data-sharing and privacy concerns that come with raw genotypes. Second, our approach is linear in the number of SNPs avoiding the computational bottleneck required to estimate the genetic relationship matrix. Conceptually, our approach is very similar to that taken by LD score regression. Indeed, the diagonal of the LD matrix product in one population are exactly the LD-scores ($\Sigma_{Aii}^2 = l_i$). One could ignore our likelihood-based approach and define cross-population scores $c_i = \sum_m r_{1im} r_{2im}$ in order to exploit the linear relationship $\mathbb{E}[Z_{1i} Z_{2i}] = \frac{\sqrt{N_1 N_2}}{M} \rho_{gi} \sqrt{h_1^2 h_2^2} c_i$ (a similar approach can be taken for the genetic effect correlation). Since LD-score regression has been successfully used to compute the genetic correlation of two phenotypes in a single population, this derivation can be viewed as an extension of LD-score regression to one phenotype in two different populations. The main difference in our approach is choosing maximum likelihood rather than regression in order to fit the model. A comparison of our method to the ldsc software shows they perform similarly as heritability estimators (Figure 3.8).

Of course, our method is not without drawbacks. First, it requires a large sample size and large number of SNPs to achieve standard errors low enough to generate accurate estimates. Until recently large sample GWAS have been rare in non-European populations, though they are becoming more common. Similarly, reference panel quality may suffer in non-European

populations and this may impact downstream analysis [69]. Second, it is limited to analyzing relatively common SNPs, both because having an accurate disease model is important for the analysis of rare variants and because effect size and correlation coefficient estimates have a high standard error at rare SNPs [7]. Third, our analysis is currently limited to SNPs that are present in both populations. Indeed it is currently unclear how best to handle population-specific variants in this framework. Fourth, our estimator of ρ is bounded between -1 and 1. This may induce bias when the true value is close to the boundary and the sample size is small. Finally, admixed populations induce very long-range LD that is not accounted for in our approach and we are therefore limited to un-admixed populations [7].

Our analysis leaves open several avenues for future work. We approximately maximize the likelihood of an $M \times M$ multivariate normal distribution via a method that uses only the diagonal elements of each block, discarding covariance information between Z-scores. A better approximation may lower the standard error of the estimator, facilitating an analysis of the genetic correlation of functional categories, pathways and genetic regions. We would also like to extend our analysis to include population specific variants as well as variants at frequencies between 1-5% or lower than 1%. Our simulations indicate that having an accurate disease model is important for determining the difference between the genetic effect and genetic impact correlation when rare variants are included. Maximum likelihood approaches are well suited to different genetic architectures, for example one could explicitly model the relationship between allele frequency and effect size $2f(1-f) \propto \beta^{\alpha}$, where $\alpha = -1$ corresponds to the inverse assumption and $\alpha = 0$ corresponds to the independence assumption. One could go even further and incorporate population divergence into the prior distribution of the effects sizes, modeling the relationship between effect sizes as a function of F_{st}, which may reveal important biological insights. In future work, we hope to model these effects while incorporating additional sources of information such as the effect of selection.

Chapter 4

Allele-specific transcript abundance estimation

4.1 Introduction

Humans are diploid organisms, carrying two copies of each gene. If an individual is heterozygous for an exonic variant, this can be leveraged to separately infer the abundances of the transcripts with and without the variant. These *allele-specific* (AS) estimates of transcript abundances are of substantial interest. AS differences in abundance in an individual are biologically interesting in their own right, and when combined with parental genotype information they can be used to detect genomic imprinting. AS estimates can also be used downstream in analysis of transcript counts to infer e.g. effects of *cis*-regulatory variants.

Unfortunately, AS estimates are under-utilized in practice due to numerous technical challenges. Chief among these is that of mapping bias. Reads from the haplotype containing the alternative variant contain at least one mismatch *a priori*, and therefore have a lower probability of mapping accurately than reads from the reference haplotype when using traditional alignment methods. While variant-aware aligners can reduce this bias somewhat, the most accurate method is to either i) align to a personalized reference transcriptome or ii) filter SNPs that show mapping bias in simulation [17]. Alignment to a personalized reference transcriptome eliminates mapping bias by removing the concept of a reference allele altogether. This method involves creating a reference transcriptome containing two copies of each gene, where each copy contains the variants from one of the haplotypes of an individual. While complete assembly of a personalized reference genome is not possible due to structural variants with incomplete positional information, the overwhelming majority of exonic genetic variation consists of SNPs and short indels which can be accommodated in this fashion.

Creating personalized reference transcriptomes requires knowing which variants lie on the same haplotypes. In other words, it requires (gene-level) phasing information. If high-quality haplotypes are not provided, this information is almost always acquired by statistical

phasing [25]. However, statistical phasing can be problematic in this application for several reasons. First, it adds an additional time-consuming step to an already time-consuming analysis pipeline. Second, if a switch error occurs between two variants, then reads containing both variants will map equally-well to both haplotypes and are therefore rendered un-informative for determination of allele-specific abundances. Not only are these reads crucial for that problem, but they also contain phase information and can be used to correct switch errors, particularly at rare variants [16].

In this paper, we introduce `ursa`, a tool for quantifying AS abundances from RNA-seq data and individual genotypes. `ursa` leverages the pseudoalignment of `kallisto` and phasing information in the RNA-seq reads to quantify the abundances without time-consuming read alignment or genotype phasing steps. When haplotypes are known, `ursa` creates personalized reference transcriptomes, estimates abundances with `kallisto` and post-processes it's output. When haplotypes are unknown, `ursa` creates personalized references corresponding to all possible phasings of each gene, and relies on phasing-informative reads to cause the EM algorithm of `kallisto` to converge to the correct AS abundances.

We show that this approach eliminates mapping bias on-average and yields accurate estimates of the AS abundances even without haplotype information, as compared to the only other known tool for estimating AS abundances from RNA-seq and genotype data, `allelecounter` [17].

4.2 Results

We first investigated the feasibility of building a personalized reference transcriptome containing all possible phasings of each gene. It has previously been shown that the number of hets per person per gene is low ([17] Figure 2D). We replicate that finding (Figure 4.1) and observe that 95% of genes have fewer than 8 het SNPs. By filtering gene-person pairs containing more than 8 het SNPs, and building one transcript per possible phasing of each gene, we increase the size of the reference transcriptome by 5.5× on average across all individuals in the gEUVADIS dataset. Therefore we conclude that, contrary to intuition, building a personalized transcriptome containing all possible phasings of each gene is not computationally difficult.

Next, we used simulated RNA-seq data to evaluate the performance of `ursa` with and without haplotype information. We compared `ursa` to the only other known method for estimating allele-specific counts using just genotype and RNA-seq data, `allelecounter`. We simulated haplotype-specific "ground-truth" count matrices for each gene in each person from negative-binomial distributions with parameters learned from real data, and simulated reads corresponding to these count data using an error model learned from real data. See Section 4.3 and Figure 4.6 for details. The resulting mean simulated reference and alternate allele counts were 137.9 and 137.5, with standard deviations of 1249.43 and 1031.34 respectively. Figure 4.2 shows a kernel density estimate (KDE) of the distribution of the reference and alt counts.

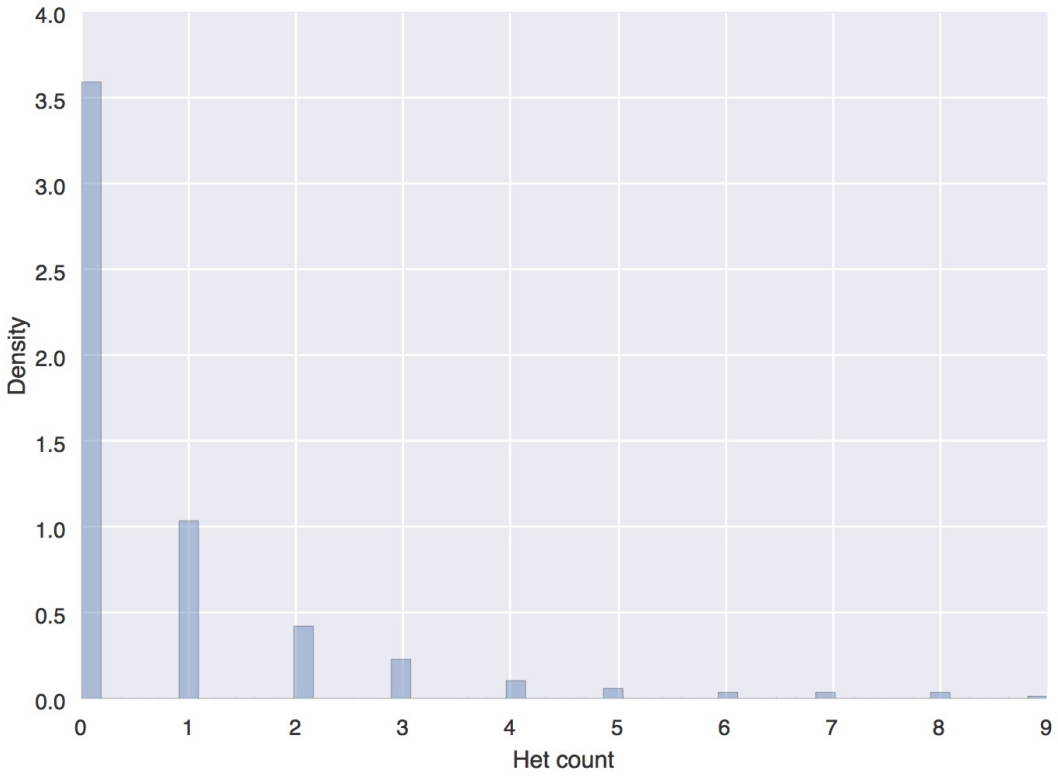

Figure 4.1: Distribution of heterozygous SNPs in gEUVADIS individuals. We replicate the finding of Castel et al [17] that most genes contain few hets, and observe that 95% of genes have 8 or fewer heterozygous SNPs.

Table 4.1 shows performance indicators for i) `ursa` with haplotype information, ii) `ursa` without haplotype information and iii) `allelecounter`. With and without haplotype information, `ursa` provides unbiased estimates of the reference and alternate allele counts and shows extremely strong correlation with the true counts, while `allelecounter` shows a strong downward bias in its estimates and only modest correlation with the true counts. In Figure 4.3 we plot the true versus estimated counts for all methods considered, along with a line of best fit for a linear regression of the estimated counts on the true counts. Notice that while `ursa` still shows strong correlation with the true counts without haplotype information, there is clearly additional error. Finally, in Figure 4.4 we compare the estimated counts from `ursa` with and without haplotype information. Clearly, there is high concordance between the two estimators, with a correlation coefficient of $\rho = 0.995$ for the reference count and a correlation coefficient of $\rho = 0.994$ for the alternate count.

The performance difference between `ursa` with and without haplotypes is more easily visualized by looking at the distribution of the count errors of each case. This is presented

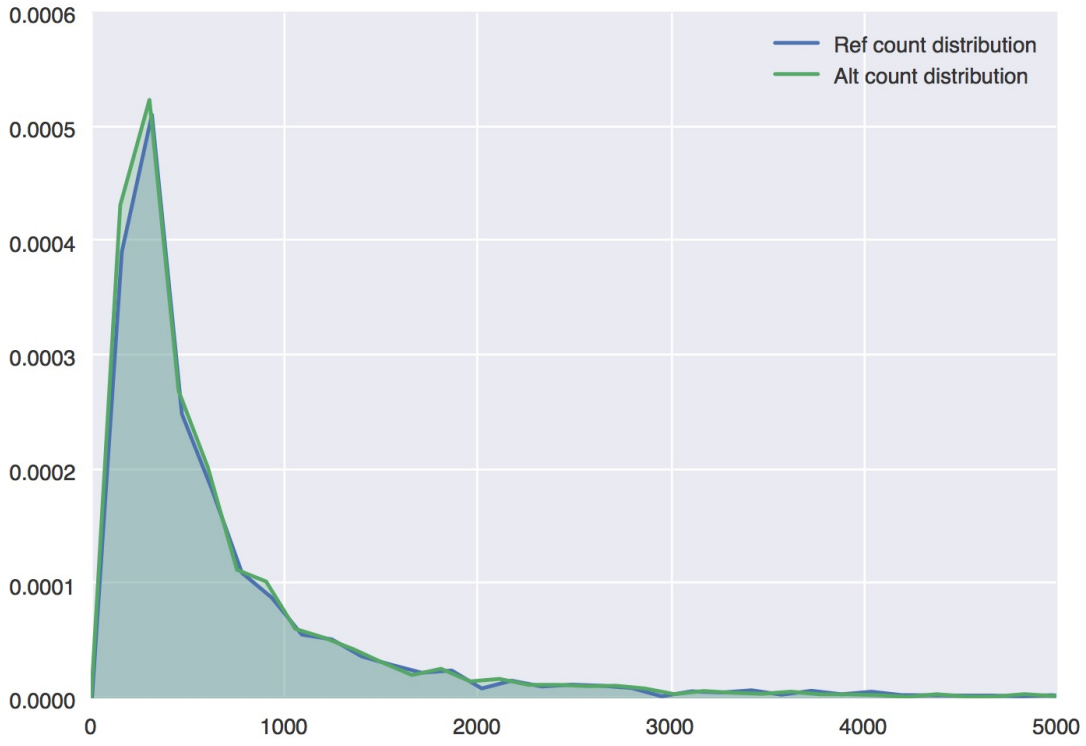

Figure 4.2: KDE of the distribution of the "ground-truth" counts generated by our simulation. Due to the long tail of the distribution, we limit the x-axis for improved visualization.

Table 4.1: Performance of `ursa` with and without haplotype information as compared to `allelecounter` for estimation of reference and alternate allele counts in simulation. ME is the mean error, RMSE is the square root of the mean squared error, and Cor is the correlation of the estimated counts to the simulated counts.

	Reference allele				Alternate allele			
Estimator	ME	SE_{ME}	RMSE	Cor	ME	SE_{ME}	RMSE	Cor
`ursa` w/ haps	-0.451	0.284	80.241	0.998	-0.513	0.289	81.723	0.997
`ursa` w/o haps	-0.564	0.470	133.021	0.994	-0.304	0.474	134.017	0.992
`allelecounter`	-120.302	4.269	1212.828	0.474	-121.018	3.493	995.225	0.511

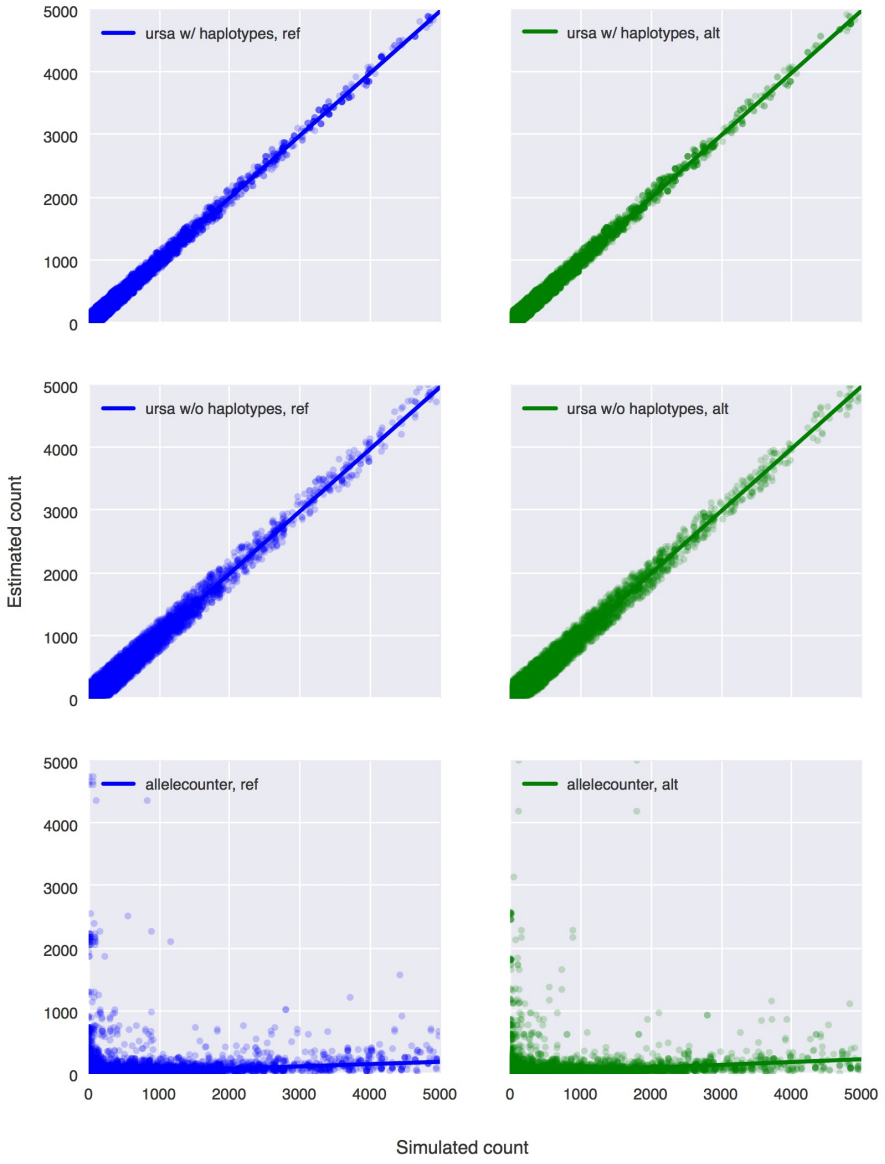

Figure 4.3: True (x-axis) versus estimated (y-axis) counts for reference (blue) and alternate (green) alleles using (top) `ursa` with haplotypes, (middle) `ursa` without haplotypes and (bottom) `allelecounter`. In each case, we overlay the line of best fit from a linear regression of the estimated counts on the true counts.

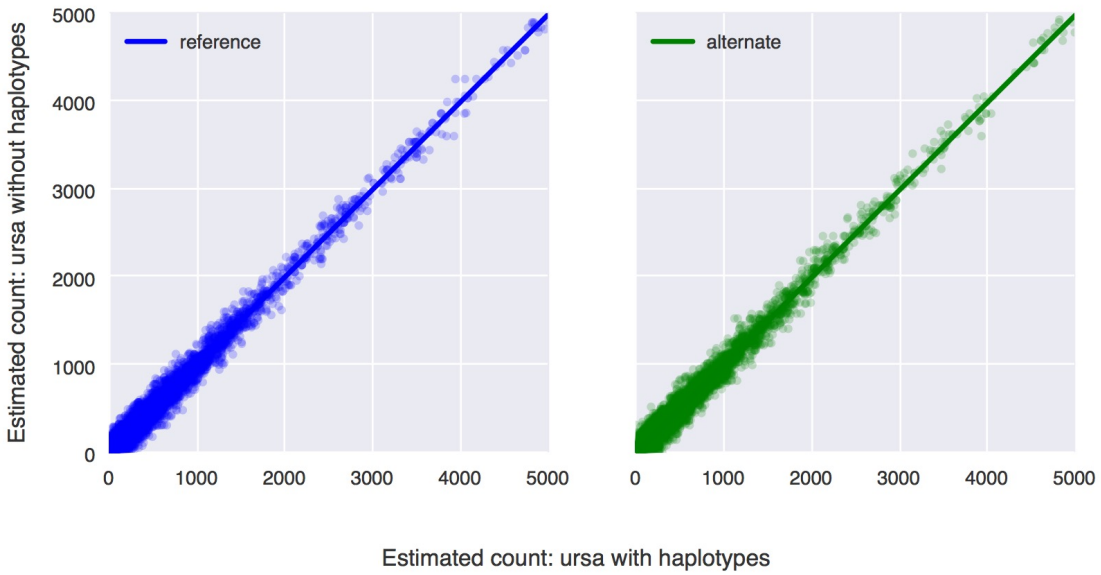

Figure 4.4: Comparison of estimates of reference (blue) and alternate (green) counts using `ursa` with (x-axis) and without (y-axis) haplotype information. The estimates are highly concordant, with a correlation coefficient of $\rho = 0.995$ for the reference count and a correlation coefficient of $\rho = 0.994$ for the alternate count.

in Figure 4.5, which shows a violin plot of the errors in these two cases. While `ursa` provides accurate estimates of the reference and alternate counts when haplotypes are not provided, the difference in performance here clearly indicates that `ursa` is unable to completely resolve haplotype phasing in all cases.

4.3 Methods

Generation of personalized and all-phasings transcriptomes

We obtained VCF files for the Phase 3 1000 genomes project individuals from the EMBL European Bioinformatics institute and filtered them to remove indels. We obtained FASTQ files for QC+ mRNA-seq reads from the European Nucleotide Archive. We obtained the hg19 human genome reference FASTA files and the hg19 GTF from the UCSC genome

Figure 4.5: Violin plot of the error in the count estimate for `ursa` with and without haplotype information.

browser. We generated 82,960 reference transcript targets from the hg19 reference and GTF using `rsem-prepare-reference`. We generated personalized reference transcriptomes using `ursa`. We define a personalized reference transcriptome as a transcriptome containing two copies of each transcript, one for each haplototype, where each copy contains every SNP that that individual has on that haplotype. If an individual has no het SNPs in a transcript, the personalized reference contains only one copy of that transcript. All phasings transcriptomes are generated by ignoring the haplotype information contained in the 1000 genomes VCF, and building a personalized reference transcriptome containing $2^{\#hets}$ targets for each transcript - one for each potential phasing. If the number of het SNPs in a transcript exceeds a user defined number (default: 8), then that transcript is left out of that individuals all-phasings transcriptome.

Simulation of RNA-seq reads

In an idealized RNA-seq experiment, where every read maps uniquely to a single transcript, the observed counts should follow a multinomial distribution [82]. This can be well-approximated by a set of independent Poisson random variables, where the variance is equal to the mean. In practice, however, the variance of the counts at each transcript is greater than the mean [77]. To model this over-dispersion, many authors choose to use a negative binomial (NB) distribution to model count data [65], and we will do the same. Similar to Pimentel et al [91], we learn the parameters of the NB distribution for each transcript from a single gender in a single population to minimize the effect of population structure, differential environment and gender on our estimates. In this work, we chose the Tuscan female (TSI-F) population of gEUVADIS. First, we quantified the transcript-level abundances of each of the TSI-F individuals using `kallisto`. Then, we used these transcript-level abundances to compute the mean and variance of the count distribution for each transcript in the TSI-F population. The mean-variance parameterization of the NB distribution can be easily converted to the more common $\mathrm{NB}(k; r, p)$ parameterization via the relationship

$$p = \frac{\sigma^2 - \mu}{\sigma^2}$$
$$r = \frac{\mu^2}{\sigma^2 - \mu}$$

where r is the number of failures in a series of $N = k + r$ bernoulli trials with success probability p.

These parameters govern the transcript counts, but are not haplotype specific. It is straightforward to observe that if $C_1, C_2 \sim \mathrm{NB}\left(k; \frac{r}{2}, p\right)$, then $C_1 + C_2 \sim \mathrm{NB}\left(k; r, p\right)$. That is, we draw counts independently for each haplotype, such that the resulting sum of the counts has the desired distribution. By simulating the counts independently, we also have the opportunity to add allele-specific effects. Specifically, if we adopt the common assumption that genetic effects are linear in log-count space [91] and we assume allele-specific effect sizes are normally distributed in log-count space, then the same effects are log-normally distributed in count space. Therefore we draw $\beta_i \sim \mathcal{N}(0, \sigma^2)$ for each SNP, so that the allele-specific count in individual j is

$$C'_{H,t,j} = \exp(\mathbb{1}[H_{ij} = 1]\beta_i) C_{H,t,j} \tag{4.1}$$

where $\mathbb{1}[H_{ij} = 1]$ is an indicator that individual j has variant i in haplotype H.

The primary sources of error and bias in RNA-seq are i) read error, ii) positional bias, and iii) sequence bias. RSEM [58] provides a thorough simulation suite that learns parameters for all of the above, as well as the fragment length distribution, from real data. First, we choose one individual from gEUVADIS to train the simulator on, in this case NA20517. Then, we quantify the transcript abundances in that individual using RSEM to learn the error model parameters. Rather than simulate reads from the human reference transcriptome,

we simulate them from personalized haplotype-specific reference transcriptomes build by `ursa` using the gEUVADIS individual haplotypes. Given the simulated haplotype-specific counts $C'_{H,t,j}$, the total number of reads for each haplotype of each individual is given by $N_{H,j} = \sum_t C'_{H,t,j}$. We therefore simulate $N_{H,j}$ paried-end reads according to the generated counts $C'_{H,t,j}$. Following this simulation, we combine the first and second reads generated by each haplotype to yield N_j total reads without haplotype of origin information. The simulation pipeline is maintained using `sakemake`. For a visualization of the simulation framework, see Figure 4.6

Quantification of AS counts

For personalized transcriptomes with or without phasing information, we quantified abundances by first building the `kallisto index` with default parameters, then quantifying with `kallisto quant` using 10 bootstrap iterations. The resulting abundances were post-processed by `ursa` to provide allele-specific reference and alt counts for each het SNP in each individual, in an output format similar to that generated by `allelecounter`. When phasing information is not provided, the allele count for a het SNP is determine by summing the counts of all possible phasings containing that SNP.

For quantification of AS counts with `allelecounter`, we set-up the pipeline described in [17] with one exception: we do not filter sites that are known to show mapping bias in simulation. This choice was made to provide a fair comparison of the ability of each method to quantify AS counts genome-wide. We used `rsem-prepare-reference --star` to build STAR alignment indices according to the ENCODE3 STAR-RSEM pipeline. We aligned reads to the STAR incides using STAR 2.5.2b, indexed the read alignment with `samtools` and marked duplicate reads using `picard`, before quantification with `allelecounter` v0.5. We set `allelecounter` parameters for minimum coverage, minimim base quality, and minimum map quality to 2, 10 and 10, respectively.

4.4 Discussion

We have presented an approach to obtaining allele-specific transcript abundance estimates that eliminates mapping bias via psuedo-alignment to personalized reference transcriptomes. When haplotype phasing information is known, we obtain extremely accurate estimates of allele-specific counts. When haplotype phasing is unknown, we pseudo-align to targets consisting of all possible phasings of each gene, and suffer little loss of accuracy in simulation. Our simulation framework is thorough, using an error model learned from real data and transcript count distributions learned from a single gender in a single population to minimize the effect of population structure, different environment and sex-specificity.

That said, differences in the accuracy with and without known haplotypes remain. There are several approaches that could be considered in future work to improve the accuracy of allele-specific estimates. First, we have made no internal modifications to the `kallisto` EM

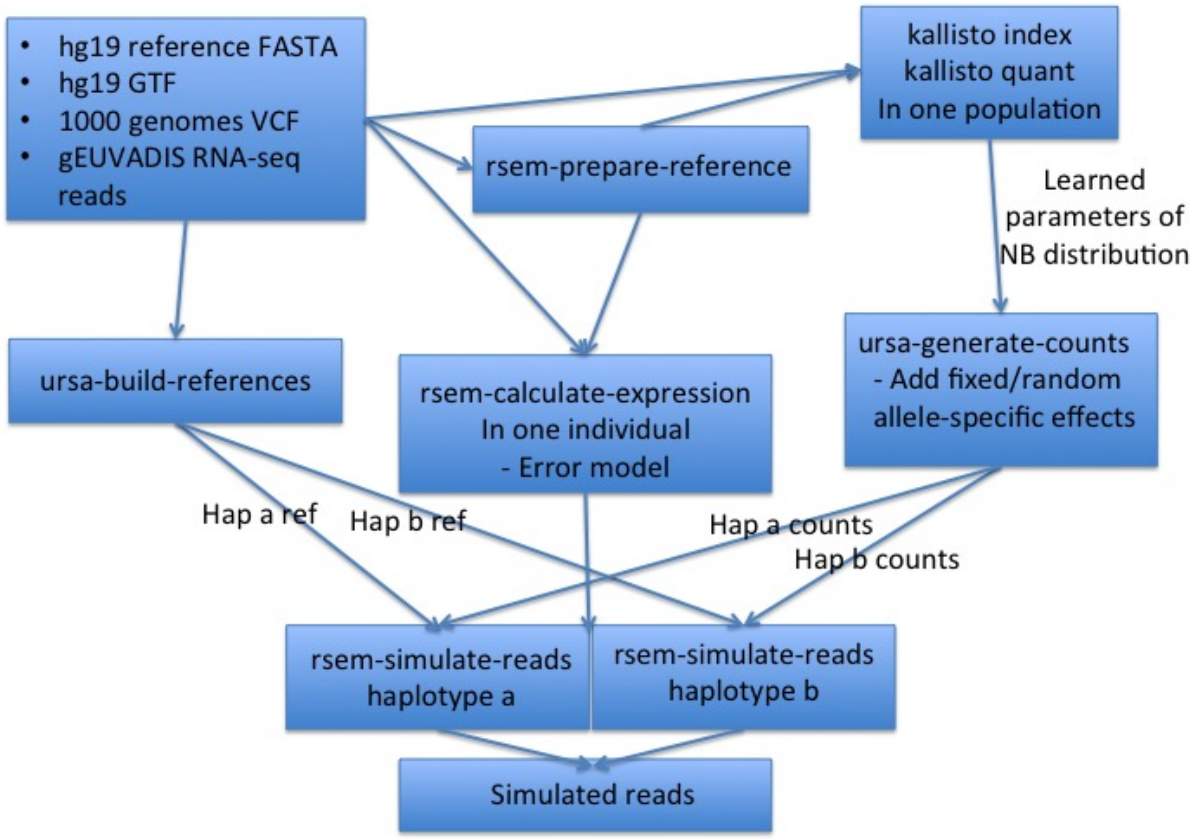

Figure 4.6: Simulation pipeline for generating personalized RNA-seq reads with allele-specific abundances. The hg19 reference FASTA and GTF are used to build the standard reference transcriptome, which is used in quantification of transcript abundances in one population with `kallisto`. These abundances are used to calculate parameters of NB distributions for the counts of each gene. These distributions are used to draw counts for the simulation, and fixed or random allele-specific effects can be added. The 1000 genomes VCFs are used by `ursa` to build personalized references. RSEM is used to build an error model for the simulator. Finally, the RSEM simulator is run independently for each haplotype with haplotype-specific counts to generate personalized RNA-seq reads, which are combined to erase haplotype of origin.

algorithm, despite the additional structure that can be exploited in this problem. While this work demonstrates that sometimes the simplest approach can yield accurate results, we may be able to improve the accuracy of estimates by, for example, constraining the likelihood to two-haplotypes and finding the pair of possible phasings that maximizes the likelihood. Specifically we can write the likelihood of the abundances as a function of the haplotypes

$$\mathcal{L}(\alpha; Y) = \sum_{h_1, h_2} \mathcal{L}(\alpha; Y, h_1, h_2)\mathbb{P}(h_1, h_2) \tag{4.2}$$

The prior distribution of the haplotypes $\mathbb{P}(h_1, h_2)$ can be learned using known methods such as SHAPEIT [25], which can compute a probability distribution over the possible phasings of a genotype. This distribution can be further used to refine the space of possible phasings that we use to determine allele-specific counts by constructing, for example, the 99% confidence set of possible phasings. Unfortunately, pre-phasing and sampling from this space will take substantial time, but may be worthwhile if it provides increased accuracy.

We compare our method to `allelecounter` because it is the only known alternative method for generating AS counts from genotypes and RNA-seq data, however we must point out that `allelecounter` is solving a slightly different problem, and that this discrepancy explains the performance issue in this application. Rather than estimate transcript abundances corresponding to the transcript containing each allele, it estimates the number of reads containing each allele. Therefore, the counts reported by `allelecounter` do not include reads that map to the transcript but do not contain the allele of interest. On the other hand, we proportionally assign these reads to the allele-specific transcripts via EM.

Note that, throughout this work, we have been careful to refer to our abundances as allele specific and not haplotype specific. Indeed, when haplotype information is provided, these are equivalent. But when haplotype information is not provided, `ursa` remains phase agnostic. By integrating the above pre-phasing approach to generate confidence sets of possible phasings, and using EM to generate a posterior distribution on the possible phasings, we may be able to obtain highly accurate haplotype specific estimates of transcript abundance.

The principal utility of this software will surely be in the improved downstream analysis of functional genomic effects. In future work, we plan on quantifying AS counts in the gEUVADIS dataset, and using them to gain insight to the structure of gene regulation. For example, AS differences in abundances might be present not just in one individual, but in many individuals in the population that share a het-SNP at a gene. Furthermore, these estimates can be used to find genetic variants associated with a difference in allele count. In general, we hope that accurate estimates of AS counts will improve power to detect genetic regulatory loci.

Chapter 5

Discussion

The genetic architecture of complex traits is a complex and fascinating topic. In the long-term, investigation in this area will simultaneously bolster our understanding of both molecular biology and evolution - topics somewhat distantly related until the recent past. From a practical perspective, complex trait genetics has numerous medical consequences: finding new drug targets, predicting side effects, predicting lifetime risk for disease and more. These findings will enable medical care to be tailored to a persons individual genome. However, discovery of these findings is limited by the statistical and computational power required to analyze massive datasets and integrate heterogeneous data types. In this work, we have made three major contributions to this problem: increased power to detect hidden associations and find missing heritability, improving our understanding of the relationship between complex trait genetics in different populations, and more accurate estimates of allele-specific counts from RNA-seq studies. There are near limitless avenues for future work in this field, but we will discuss a few specific ones in this section.

As we have discussed, the vast majority of studies of the genetics of complex traits have been conducted in European populations. Those that have not been conducted in European populations have been conducted primarily in Asian populations by Asian institutions. However, the largest amount of genetic diversity is present in Africa, where the infrastructure for conducting human genetic research is left wanting. A huge amount of medically actionable genetic variation can almost certainly be learned by more complete studies of African genetics. From a computational standpoint, we require better methods for integrating data from multiple populations. This will allow us to improve power to detect medically actionable genetic variation, and allow us to understand which discoveries can be shared across populations and which cannot. That said, as time goes on, the world is becoming increasingly admixed. An admixed individual shares genetics from multiple world populations. This is particularly apparent in America, where more than 25% of the population are African American (AA) or Hispanic/Latino. Medical genetics in admixed populations is difficult: one genetic locus could have African ancestry in 50% of your AA samples, and European ancestry in the other 50%. Methods for studying complex trait genetics in admixed populations are one of the most promising avenues for medical impact in genetics.

Much remains to be learned about the molecular path from genotype to phenotype, and very few genetic variants that appear to causally increase risk for complex disease have a well-understood biological mechanism. The amount of molecular phenotype data - such as gene expression in different tissues, chromatin accessibility and methylation status - is increasing rapidly as the cost of obtaining the data decreases. The integration of these and other data types to form a complete picture of the path from genotype to phenotype is both extremely promising and extremely challenging. In this area, better computational methods for accurately measuring the molecular phenotype of interest from the raw sequencing data are still required. For example, in this work we improved the ability to detect allele-specific transcript counts. In future work we can use these counts to find new biology. For example, we can look for genetic variants that are associated with differential expression between the two haplotypes of an individual in one population, or use these allele-specific counts to improve power to discover eQTLs. As these kinds of data become increasingly available in patient samples, accurate measurement of molecular phenotypes will enable progress on this problem.

Bibliography

[1] Nov. 2013. URL: https://liorpachter.wordpress.com/seq/.

[2] "A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease". In: *Nat Genet* 43.4 (Apr. 2011), pp. 339–344. ISSN: 1061-4036.

[3] Simon Anders and Wolfgang Huber. "Differential expression analysis for sequence count data". In: *Genome biology* 11.10 (2010), p. 1.

[4] Ya'ara Arkin et al. "EPIQ—efficient detection of SNP–SNP epistatic interactions for quantitative traits". In: *Bioinformatics* 30.12 (2014), pp. i19–i25.

[5] Nicolas L Bray et al. "Near-optimal probabilistic RNA-seq quantification". In: *Nature biotechnology* 34.5 (2016), pp. 525–527.

[6] Jason D Buenrostro et al. "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position". In: *Nature methods* 10.12 (2013), pp. 1213–1218.

[7] Brendan K Bulik-Sullivan et al. "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies". In: *Nature genetics* 47.3 (2015), pp. 291–295.

[8] Brendan Bulik-Sullivan et al. "An atlas of genetic correlations across human diseases and traits". In: *Nat Genet* 47.11 (Nov. 2015), pp. 1236–1241. ISSN: 1061-4036.

[9] MG1 Bulmer. "The effect of selection on genetic variability". In: *American Naturalist* (1971), pp. 201–211.

[10] V. L. Burt et al. "Prevalence of hypertension in the US adult population. Results from the Third National Health and Nutrition Examination Survey, 1988-1991". In: *Hypertension* 25.3 (Mar. 1995), pp. 305–313. ISSN: 0194-911X.

[11] Paul R Burton et al. "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls". In: *Nature* 447.7145 (2007), pp. 661–678.

[12] Carlos D. Bustamante, Francisco M. De La Vega, and Esteban G. Burchard. "Genomics for the world". In: *Nature* 475.7355 (July 14, 2011), pp. 163–165. ISSN: 0028-0836.

[13] Sara A Byron et al. "Translating RNA sequencing into clinical diagnostics: opportunities and challenges". In: *Nature Reviews Genetics* (2016).

[14] Teresa R. de Candia et al. "Additive Genetic Variation in Schizophrenia Risk Is Shared by Populations of African and European Descent". In: *The American Journal of Human Genetics* 93.3 (Sept. 5, 2013), pp. 463–470. ISSN: 0002-9297.

[15] Francesco Paolo Casale et al. "Efficient set tests for the genetic analysis of correlated traits". In: *Nature methods* 12.8 (2015), pp. 755–758.

[16] Stephane E Castel et al. "Rare variant phasing and haplotypic expression from RNA sequencing with phASER". In: *Nature Communications* 7 (2016).

[17] Stephane E Castel et al. "Tools and best practices for data processing in allelic expression analysis". In: *Genome biology* 16.1 (2015), p. 1.

[18] Christopher C Chang et al. "Second-generation PLINK: rising to the challenge of larger and richer datasets". In: *Gigascience* 4.1 (2015), p. 1.

[19] Man-huei Chang et al. "Racial/Ethnic Variation in the Association of Lipid-Related Genetic Variants With Blood Lipids in the US Adult Population". In: *Circulation: Cardiovascular Genetics* 4.5 (Oct. 1, 2011), pp. 523–533.

[20] Xu Chen et al. "Dominant Genetic Variation and Missing Heritability for Human Complex Traits: Insights from Twin versus Genome-wide Common SNP Models". In: *The American Journal of Human Genetics* 97.5 (Nov. 2015), pp. 708–714. ISSN: 00029297.

[21] Yoon Shin Cho et al. "Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians". In: *Nat Genet* 44.1 (Jan. 2012), pp. 67–72. ISSN: 1061-4036.

[22] 1000 Genomes Project Consortium et al. "A global reference for human genetic variation". In: *Nature* 526.7571 (2015), pp. 68–74.

[23] Marc A. Coram et al. "Leveraging Multi-ethnic Evidence for Mapping Complex Traits in Minority Populations: An Empirical Bayes Approach". In: *The American Journal of Human Genetics* 96.5 (May 2015), pp. 740–752. ISSN: 00029297.

[24] Olivia Corradin et al. "Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits". In: *Genome research* 24.1 (2014), pp. 1–13.

[25] Olivier Delaneau, Jonathan Marchini, and Jean-Franois Zagury. "A linear complexity phasing method for thousands of genomes". In: *Nature methods* 9.2 (2012), pp. 179–181.

[26] Everett R Dempster and I Michael Lerner. "Heritability of threshold characters". In: *Genetics* 35.2 (1950), p. 212.

[27] John D Eicher et al. "GRASP v2. 0: an update on the Genome-Wide Repository of Associations between SNPs and phenotypes". In: *Nucleic acids research* 43.D1 (2015), pp. D799–D804.

[28] Evan E Eichler et al. "Missing heritability and strategies for finding the underlying causes of complex disease". In: *Nature Reviews Genetics* 11.6 (2010), pp. 446–450.

[29] Michael B Elowitz et al. "Stochastic gene expression in a single cell". In: *Science* 297.5584 (2002), pp. 1183–1186.

[30] Douglas S Falconer. "The inheritance of liability to certain diseases, estimated from the incidence among relatives". In: *Annals of Human Genetics* 29.1 (1965), pp. 51–76.

[31] Douglas S Falconer, Trudy FC Mackay, and Richard Frankham. "Introduction to quantitative genetics (4th edn)". In: *Trends in Genetics* 12.7 (1996), p. 280.

[32] Jacques Fellay et al. "ITPA gene variants protect against anaemia in patients treated for chronic hepatitis C". In: *Nature* 464.7287 (2010), pp. 405–408.

[33] MD Fesinmeyer et al. "Genetic risk factors for body mass index and obesity in an ethnically diverse population: results from the Population Architecture using Genomics and Epidemiology (PAGE) Study". In: *Obesity (Silver Spring, Md.)* 21.4 (Apr. 2013), 10.1002/oby.20268. ISSN: 1930-7381. JSTOR: `{PMC}3482415`.

[34] Hilary K Finucane et al. "Partitioning heritability by functional annotation using genome wide association summary statistics". In: *Nat Genet* 47.11 (Nov. 2015), pp. 1228–1235. ISSN: 1061-4036.

[35] Ronald Aylmer Fisher. *The genetical theory of natural selection: a complete variorum edition*. Oxford University Press, 1930.

[36] Daniel J Gaffney. "Global Properties and Functional Complexity of Human Gene Regulatory Variation". In: *PLoS Genetics* 9.5 (May 2013). Ed. by Gonalo R Abecasis, e1003501. ISSN: 1553-7390. JSTOR: `{PMC}3667745`.

[37] Eric R Gamazon, Nancy J Cox, and Lea K Davis. "Structural architecture of SNP effects on complex traits". In: *The American Journal of Human Genetics* 95.5 (2014), pp. 477–489.

[38] Richard A Gibbs et al. "The international HapMap project". In: *Nature* 426.6968 (2003), pp. 789–796.

[39] Alexander Gusev et al. "Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases". In: *The American Journal of Human Genetics* 95.5 (2014), pp. 535–552.

[40] Alexander Gusev et al. "Quantifying missing heritability at known GWAS loci". In: *PLoS Genet* 9.12 (2013), e1003993.

[41] Melissa Gymrek et al. "Identifying personal genomes by surname inference". In: *Science* 339.6117 (2013), pp. 321–324.

[42] Brian J Haas, Michael C Zody, et al. "Advancing RNA-seq analysis". In: *Nature biotechnology* 28.5 (2010), p. 421.

[43] David Haig. "Does heritability hide in epistasis between linked SNPs?" In: *European Journal of Human Genetics* 19.2 (2011), p. 123.

[44] Buhm Han, Hyun Min Kang, and Eleazar Eskin. "Rapid and accurate multiple testing correction and power estimation for millions of correlated markers". In: *PLoS Genet* 5.4 (2009), e1000456.

[45] Gibran Hemani, Sara Knott, and Chris Haley. "An evolutionary perspective on epistasis and the missing heritability". In: *PLoS Genet* 9.2 (2013), e1003295.

[46] Farhad Hormozdiari et al. "Identification of causal genes for complex traits". In: *Bioinformatics* 31.12 (June 15, 2015), pp. i206–i213. ISSN: 1367-4803, 1460-2059.

[47] F. Hormozdiari et al. "Identifying Causal Variants at Loci with Multiple Signals of Association". In: *Genetics* 198.2 (Oct. 1, 2014), pp. 497–508. ISSN: 0016-6731.

[48] Richard Howey and Heather J Cordell. "Imputation Without Doing Imputation: A New Method for the Detection of Non-Genotyped Causal Variants". In: *Genetic epidemiology* 38.3 (2014), pp. 173–190.

[49] International Schizophrenia Consortium. "Common polygenic variation contributes to risk of schizophrenia that overlaps with bipolar disorder". In: *Nature* 460.7256 (Aug. 6, 2009), pp. 748–752. ISSN: 0028-0836. JSTOR: `{PMC}3912837`.

[50] Iuliana Ionita-Laza et al. "Sequence kernel association tests for the combined effect of rare and common variants". In: *The American Journal of Human Genetics* 92.6 (2013), pp. 841–853.

[51] David S Johnson et al. "Genome-wide mapping of in vivo protein-DNA interactions". In: *Science* 316.5830 (2007), pp. 1497–1502.

[52] Gleb Kichaev et al. "Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies". In: *PLoS Genetics* 10.10 (Oct. 30, 2014). Ed. by Anna Di Rienzo, e1004722. ISSN: 1553-7404.

[53] Tuuli Lappalainen et al. "Epistatic selection between coding and regulatory variation in human evolution and disease". In: *The American Journal of Human Genetics* 89.3 (2011), pp. 459–463.

[54] Charity W Law et al. "Voom: precision weights unlock linear model analysis tools for RNA-seq read counts". In: *Genome biology* 15.2 (2014), p. 1.

[55] S. H. Lee et al. "Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood". In: *Bioinformatics* 28.19 (Oct. 1, 2012), pp. 2540–2542. ISSN: 1367-4803, 1460-2059.

[56] Sang Hong Lee et al. "Estimating missing heritability for disease from genome-wide association studies". In: *The American Journal of Human Genetics* 88.3 (2011), pp. 294–305.

[57] Seunggeun Lee et al. "General Framework for Meta-analysis of Rare Variants in Sequencing Association Studies". In: *The American Journal of Human Genetics* 93.1 (July 2013), pp. 42–53. ISSN: 00029297.

[58] Bo Li and Colin N Dewey. "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome". In: *BMC bioinformatics* 12.1 (2011), p. 1.

[59] DY Lin. "An efficient Monte Carlo approach to assessing statistical significance in genomic studies". In: *Bioinformatics* 21.6 (2005), pp. 781–787.

[60] Jennifer Listgarten et al. "A powerful and efficient set test for genetic markers that handles confounders". In: *Bioinformatics* 29.12 (2013), pp. 1526–1533.

[61] Jennifer Listgarten et al. "Improved linear mixed models for genome-wide association studies". In: *Nature methods* 9.6 (2012), pp. 525–526.

[62] Jimmy Z Liu et al. "Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis". In: *Nature genetics* 44.10 (2012), pp. 1137–1141.

[63] John Locke. *An essay concerning human understanding*. 1841.

[64] Po-Ru Loh et al. "Efficient Bayesian mixed-model analysis increases association power in large cohorts". In: *Nature genetics* 47.3 (2015), pp. 284–290.

[65] Michael I Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome biology* 15.12 (2014), p. 1.

[66] Julius B Lucks et al. "Multiplexed RNA structure characterization with selective 2-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq)". In: *Proceedings of the National Academy of Sciences* 108.27 (2011), pp. 11063–11068.

[67] Michael Lynch, Bruce Walsh, et al. *Genetics and analysis of quantitative traits*. Vol. 1. Sinauer Sunderland, MA, 1998.

[68] Teri A Manolio et al. "Finding the missing heritability of complex diseases". In: *Nature* 461.7265 (2009), pp. 747–753.

[69] Jonathan Marchini and Bryan Howie. "Genotype imputation for genome-wide association studies". In: *Nat Rev Genet* 11.7 (July 2010), pp. 499–511. ISSN: 1471-0056.

[70] M. E. Marenberg et al. "Genetic susceptibility to death from coronary heart disease in a study of twins". In: *The New England Journal of Medicine* 330.15 (Apr. 14, 1994), pp. 1041–1046. ISSN: 0028-4793.

[71] John C Marioni et al. "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays". In: *Genome research* 18.9 (2008), pp. 1509–1517.

[72] Alexander Meissner et al. "Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis". In: *Nucleic acids research* 33.18 (2005), pp. 5868–5877.

[73] Tarjei S Mikkelsen et al. "Genome-wide maps of chromatin state in pluripotent and lineage-committed cells". In: *Nature* 448.7153 (2007), pp. 553–560.

[74] Andrew P Morris. "Transethnic Meta-Analysis of Genomewide Association Studies". In: *Genetic Epidemiology* 35.8 (Dec. 2011), pp. 809–822. ISSN: 0741-0395. JSTOR: `{PMC}3460225`.

[75] Andrew P Morris et al. "Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes". In: *Nature genetics* 44.9 (Sept. 2012), pp. 981–990. ISSN: 1061-4036. JSTOR: `{PMC}3442244`.

[76] Ali Mortazavi et al. "Mapping and quantifying mammalian transcriptomes by RNA-Seq". In: *Nature methods* 5.7 (2008), pp. 621–628.

[77] Ugrappa Nagalakshmi et al. "The transcriptional landscape of the yeast genome defined by RNA sequencing". In: *Science* 320.5881 (2008), pp. 1344–1349.

[78] Matthew R Nelson et al. "An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people". In: *Science* 337.6090 (2012), pp. 100–104.

[79] J. J. Nora et al. "Genetic–epidemiologic study of early-onset ischemic heart disease". In: *Circulation* 61.3 (Mar. 1980), pp. 503–508. ISSN: 0009-7322.

[80] Yukinori Okada et al. "Genetics of rheumatoid arthritis contributes to biology and drug discovery". In: *Nature* 506.7488 (Feb. 20, 2014), pp. 376–381. ISSN: 0028-0836. JSTOR: `{PMC}3944098`.

[81] Alicia Oshlack, Mark D Robinson, and Matthew D Young. "From RNA-seq reads to differential expression results". In: *Genome biology* 11.12 (2010), p. 1.

[82] Lior Pachter. "Models for transcript quantification from RNA-Seq". In: *arXiv preprint arXiv:1104.3889* (2011).

[83] Luigi Palla and Frank Dudbridge. "A Fast Method that Uses Polygenic Scores to Estimate the Variance Explained by Genome-wide Marker Panels and the Proportion of Variants Affecting a Trait". In: *The American Journal of Human Genetics* 97.2 (Aug. 2015), pp. 250–259. ISSN: 00029297.

[84] Danny S. Park et al. "Adapt-Mix: learning local genetic correlation structure improves summary statistics-based analyses". In: *Bioinformatics* 31.12 (June 15, 2015), pp. i181–i189. ISSN: 1367-4803, 1460-2059.

[85] Bogdan Paaniuc, Noah Zaitlen, and Eran Halperin. "Accurate estimation of expression levels of homologous genes in RNA-seq experiments". In: *Journal of Computational Biology* 18.3 (2011), pp. 459–468.

[86] B. Pasaniuc et al. "Fast and accurate imputation of summary statistics enhances evidence of functional enrichment". In: *Bioinformatics* 30.20 (Oct. 15, 2014), pp. 2906–2914. ISSN: 1367-4803, 1460-2059.

[87] Rob Patro, Stephen M Mount, and Carl Kingsford. "Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms". In: *Nature biotechnology* 32.5 (2014), pp. 462–464.

[88] Nikolaos A Patsopoulos et al. "Fine-mapping the genetic association of the major histocompatibility complex in multiple sclerosis: HLA and non-HLA effects". In: *PLoS Genet* 9.11 (2013), e1003926.

[89] Itsik Pe'er et al. "Estimation of the multiple testing burden for genomewide association studies of nearly all common variants". In: *Genetic epidemiology* 32.4 (2008), pp. 381–385.

[90] Joseph K Pickrell. "Joint analysis of functional genomic data and genome-wide association studies of 18 human traits". In: *The American Journal of Human Genetics* 94.4 (2014), pp. 559–573.

[91] Harold J Pimentel et al. "Differential analysis of RNA-Seq incorporating quantification uncertainty". In: *bioRxiv* (2016), p. 058164.

[92] Snehit Prabhu and Itsik Pe'er. "Ultrafast genome-wide scan for SNP–SNP interactions in common complex disease". In: *Genome research* 22.11 (2012), pp. 2230–2240.

[93] Alkes L. Price et al. "Single-Tissue and Cross-Tissue Heritability of Gene Expression Via Identity-by-Descent in Related or Unrelated Individuals". In: *PLoS Genetics* 7.2 (Feb. 24, 2011). Ed. by Greg Gibson, e1001317. ISSN: 1553-7404.

[94] Adam Roberts and Lior Pachter. "Streaming fragment assignment for real-time analysis of sequencing experiments". In: *Nature methods* 10.1 (2013), pp. 71–73.

[95] Adam Roberts et al. "Improving RNA-Seq expression estimates by correcting for fragment bias". In: *Genome biology* 12.3 (2011), p. 1.

[96] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". In: *Bioinformatics* 26.1 (2010), pp. 139–140.

[97] Matthew R Robinson et al. "Population genetic differentiation of height and body mass index across Europe". In: *Nature Genetics* 47.11 (Sept. 14, 2015), pp. 1357–1362. ISSN: 1061-4036, 1546-1718.

[98] SR Seaman and B Mller-Myhsok. "Rapid simulation of P values for product methods and multiple-testing adjustment in association studies". In: *The American Journal of Human Genetics* 76.3 (2005), pp. 399–408.

[99] Huwenbo Shi, Gleb Kichaev, and Bogdan Pasaniuc. "Contrasting the genetic architecture of 30 complex traits from summary association data". In: *bioRxiv* (2016), p. 035907.

[100] Thomas P Slavin et al. "Two-marker association tests yield new disease associations for coronary artery disease and hypertension". In: *Human genetics* 130.6 (2011), pp. 725–733.

[101] Hon-Cheong So, Miaoxin Li, and Pak C. Sham. "Uncovering the total heritability explained by all true susceptibility variants in a genome-wide association study". In: *Genetic Epidemiology* (2011), n/a–n/a. ISSN: 07410395.

[102] Doug Speed et al. "Improved heritability estimation from genome-wide SNPs". In: *The American Journal of Human Genetics* 91.6 (2012), pp. 1011–1021.

[103] Eli A Stahl et al. "Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis". In: *Nature Genetics* 44.5 (Mar. 25, 2012), pp. 483–489. ISSN: 1061-4036, 1546-1718.

[104] Barbara E Stranger et al. "Patterns of cis regulatory variation in diverse human populations". In: *PLoS Genet* 8.4 (2012), e1002639.

[105] Zhan Su, Jonathan Marchini, and Peter Donnelly. "HAPGEN2: simulation of multiple disease SNPs". In: *Bioinformatics* 27.16 (Aug. 15, 2011), pp. 2304–2305. ISSN: 1367-4803. JSTOR: `{PMC}3150040`.

[106] Peter A C 't Hoen et al. "Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories". In: *Nature Biotechnology* 31.11 (Sept. 15, 2013), pp. 1015–1022. ISSN: 1087-0156, 1546-1696.

[107] Jacob A Tennessen et al. "Evolution and functional impact of rare coding variation from deep sequencing of human exomes". In: *science* 337.6090 (2012), pp. 64–69.

[108] Cole Trapnell and Steven L Salzberg. "How to map billions of short reads onto genomes". In: *Nature biotechnology* 27.5 (2009), p. 455.

[109] Cole Trapnell et al. "Differential analysis of gene regulation at transcript resolution with RNA-seq". In: *Nature biotechnology* 31.1 (2013), pp. 46–53.

[110] Cole Trapnell et al. "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation". In: *Nature biotechnology* 28.5 (2010), pp. 511–515.

[111] Gosia Trynka et al. "Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease". In: *Nature genetics* 43.12 (2011), pp. 1193–1201.

[112] Miriam S Udler et al. "FGFR2 variants and breast cancer risk: fine-scale mapping using African American studies and analysis of chromatin conformation". In: *Human molecular genetics* 18.9 (2009), pp. 1692–1703.

[113] Shashaank Vattikuti, Juen Guo, and Carson C. Chow. "Heritability and Genetic Correlations Explained by Common SNPs for Metabolic Syndrome Traits". In: *PLoS Genetics* 8.3 (Mar. 29, 2012). Ed. by Peter M. Visscher, e1002637. ISSN: 1553-7404.

[114] Peter M Visscher, William G Hill, and Naomi R Wray. "Heritability in the genomics era—concepts and misconceptions". In: *Nature Reviews Genetics* 9.4 (2008), pp. 255–266.

[115] Peter M Visscher et al. "Five years of GWAS discovery". In: *The American Journal of Human Genetics* 90.1 (2012), pp. 7–24.

[116] Zhanyong Wang et al. "Gene–Gene Interactions Detection Using a Two-stage Model". In: *Journal of Computational Biology* 22.6 (2015), pp. 563–576.

[117] Zhong Wang, Mark Gerstein, and Michael Snyder. "RNA-Seq: a revolutionary tool for transcriptomics". In: *Nature reviews genetics* 10.1 (2009), pp. 57–63.

[118] Kevin M Waters et al. "Generalizability of Associations from Prostate Cancer GWAS in Multiple Populations". In: *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 18.4 (Apr. 2009), pp. 1285–1289. ISSN: 1055-9965. JSTOR: `{PMC}2917607`.

[119] Danielle Welter et al. "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations". In: *Nucleic acids research* 42.D1 (2014), pp. D1001–D1006.

[120] Andrew R Wood et al. "Allelic heterogeneity and more detailed analyses of known loci explain additional phenotypic variation and reveal complex patterns of association". In: *Human molecular genetics* 20.20 (2011), pp. 4082–4092.

[121] CF Jeff Wu. "On the convergence properties of the EM algorithm". In: *The Annals of statistics* (1983), pp. 95–103.

[122] Yi Xing et al. "An expectation-maximization algorithm for probabilistic reconstructions of full-length isoforms from splice graphs". In: *Nucleic acids research* 34.10 (2006), pp. 3150–3160.

[123] Zheng Xu et al. "DISSCO: direct imputation of summary statistics allowing covariates". In: *Bioinformatics* 31.15 (Aug. 1, 2015), pp. 2434–2442. ISSN: 1367-4803, 1460-2059.

[124] Jian Yang et al. "Common SNPs explain a large proportion of the heritability for human height". In: *Nature genetics* 42.7 (2010), pp. 565–569.

[125] Jian Yang et al. "Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits". In: *Nature genetics* 44.4 (2012), pp. 369–375.

[126] Jian Yang et al. "GCTA: a tool for genome-wide complex trait analysis". In: *The American Journal of Human Genetics* 88.1 (2011), pp. 76–82.

[127] Jian Yang et al. "Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index". In: *Nature Genetics* 47.10 (Aug. 31, 2015), pp. 1114–1120. ISSN: 1061-4036, 1546-1718.

[128] Noah Zaitlen, H Min Kang, and Eleazar Eskin. "Linkage effects and analysis of finite sample errors in the HapMap". In: *Human heredity* 68.2 (2009), pp. 73–86.

[129] Noah Zaitlen et al. "Analysis of case–control association studies with known risk variants". In: *Bioinformatics* 28.13 (2012), pp. 1729–1737.

[130] Noah Zaitlen et al. "Informed conditioning on clinical covariates increases power in case-control association studies". In: *PLoS Genet* 8.11 (2012), e1003032.

[131] Xiang Zhou and Matthew Stephens. "Genome-wide efficient mixed-model analysis for association studies". In: *Nature genetics* 44.7 (2012), pp. 821–824.

[132] Zhihong Zhu et al. "Dominance Genetic Variation Contributes Little to the Missing Heritability for Human Complex Traits". In: *The American Journal of Human Genetics* 96.3 (Mar. 5, 2015), pp. 377–385. ISSN: 0002-9297.

[133] Hui Zou and Trevor Hastie. "Regularization and variable selection via the elastic net". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2 (2005), pp. 301–320.

[134] Or Zuk et al. "The mystery of missing heritability: Genetic interactions create phantom heritability". In: *Proceedings of the National Academy of Sciences* 109.4 (2012), pp. 1193–1198.

[135] Lingjun Zuo et al. "A Novel, Functional and Replicable Risk Gene Region for Alcohol Dependence Identified by Genome-Wide Association Study". In: *PLoS ONE* 6.11 (2011). Ed. by Shree Ram Singh, e26726. ISSN: 1932-6203. JSTOR: `{PMC}3210123`.